A

Abbott, Edith (1876–1957)

P. Kerr

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

Abbott's work there involved her directly in action for the protection and education of juveniles and immigrants, for improvements in housing, and for the reform of correctional institutions. She also worked towards women's suffrage, the 10-h law to protect women in employment, and the admission of women into trades unions. In the 1930s she was to become a staunch advocate of social insurance measures and the welfare state. Although sympathetic to the New Deal, she felt it to be entirely inadequate when it came to welfare policies.

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

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

Macmillan Publishers Ltd (ed.), *The New Palgrave Dictionary of Economics*,
https://doi.org/10.1057/978-1-349-95189-5

Selected Works

1905. Wages of unskilled labour in the United States, 1850–1900. *Journal of Political Economy* 13: 321–67.
1906. Industrial employment of women in the United States. *Journal of Political Economy* 14: 461–501.
1908. Study of early history of child labour in America. *American Journal of Sociology* 14: 15–37.
1910. *Women in industry: A study in American economic history.* London: Appleton & Co; last reprinted in 1970.
1915. A forgotten minimum wage bill. *Life and Labour* 5: 13–16.

Abramovitz, Moses (1912–2000)

Richard A. Easterlin

Keywords
Abramovitz, M.; Aggregate demand theory; Business cycles; Economic growth in the very long run; Inventories; Kuznets cycles

JEL Classifications
B31

Born in Brooklyn, New York, Abramovitz was educated at Harvard (AB, 1932) and Columbia (Ph.D., 1939). He held faculty appointments at Columbia (l940–2, 1946–8) and Stanford University (1948–77) and was a member of the research staff of the National Bureau of Economic Research from 1938 to 1969. From 1942 to 1946 he worked as an economist for several organizations within the United States government. He was elected president of the American Economic Association in 1979–80.

Abramovitz's work, which was particularly influenced by Wesley C. Mitchell and Simon Kuznets, centres on the study of long-term economic growth and fluctuations in industrialized market economies. His first major contribution was an empirical study of business inventories that demonstrated the importance of inventory change in the shorter swings of the business cycle, and showed how the classification of inventories by stage of processing aided in the explanation of their behaviour (Abramovitz 1950). From this, Abramovitz went on to the study of longer-term fluctuations, Kuznets cycles of 15 to 20 years duration, and formulated the most widely accepted interpretation of these cycles. Using Keynesian aggregate demand theory, Abramovitz developed a model linking Kuznets cycles to long swings in building cycles and demographic variables, and to shorter-term business cycles (Abramovitz 1959a, 1961, 1964, 1968).

Contemporaneously with his work on fluctuations, Abramovitz made important contributions to long-term economic growth. He was one of the first to demonstrate that only a small share of long-term output growth in the United States was explained by factor inputs (Abramovitz 1956). He documented and analysed the increasing role of government during long-term economic growth (Abramovitz 1957, 1981) and directed and coordinated a comparative study of the post-war economic growth of a number of industrialized market nations (Abramovitz 1979b, 1986). Finally, he challenged in characteristically perceptive fashion the facile linkage made by many economists between economic growth and improving human welfare (Abramovitz 1959b, 1979a, 1982).

Selected Works

1950. *Inventories and business cycles.* New York: NBER.
1956. Resource and output trends in the United States since 1870. *American Economic Review, Papers and Proceedings* 46(2): 5–23.
1957. (With V. Eliasberg.) *The growth of public employment in Great Britain.* Princeton: Princeton University Press.
1959a. Long swings in U.S. economic growth. Statement presented to joint economic committee of the congress. Hearings before joint

economic committee of the congress of the U.S. on *Employment, Growth and Price Levels,* Part 2, 11–66, 10 April.
1959b. The welfare interpretation of secular trends in national income and production. In *The allocation of economic resources*: *Essays in honor of Bernard F. Haley*, ed. M. Abramovitz et al. Stanford: Stanford University Press.
1961. The nature and significance of Kuznets cycles. *Economic Development and Cultural Change* 9: 225–248.
1964. *Evidence of long swings in aggregate construction since the civil war.* Occasional paper no. 90. New York: NBER.
1968. The passing of the Kuznets cycle. *Economica* 349–367.
1979a. Economic growth and its discontents. In *Economics and human welfare*: *Essays in honor of Tibor Scitovsky,* ed. M. Boskin. New York: Academic Press.
1979b. Rapid growth potential and its realization: The experience of capitalist economies in the postwar period. In *Economic growth and resources. Proceedings of the fifth world congress of the international economic association,* vol. 1. London/New York: Macmillan.
1981. Welfare quandaries and productivity concerns. Presidential address to the American economic association. *American Economic Review* 71: 1–17.
1982. The retreat from economic advance. In *Progress and its discontents,* ed. G.A. Almond, M. Chodorow, and R.H. Pearce. Berkeley: University of California Press.
1986. Catching up, forging ahead and falling behind. *Journal of Economic History* 46: 385–406.

Absentee

F. Y. Edgeworth

An absentee may be variously defined (1) as a landed proprietor who resides away from his estate, or (2) from his country; or more generally (3) any unproductive consumer who lives out of the country from which he derives his income.

Examples of these species are (1) a seigneur under the *ancien régime* living in Paris at a distance from his estates; (2) an Irish landlord resident abroad; (3) an Anglo-Indian ex-official resident in England and drawing a pension from India. In writing briefly on the evils of absenteeism it is difficult to use general terms appropriate to all the definitions; but considerations primarily relating to some one definition may easily be adapted to another by the reader.

It is useful to consider separately the effects of the absentee proprietor's consumption upon the wealth of his countrymen; and the moral, as well as economical effects of other circumstances.

I. The more abstract question turns upon the fact that the income of an absentee is mostly remitted by means of exports. 'The tribute, subsidy, or remittance is always in goods . . . unless the country possesses mines of the precious metals' (Mill). So far as the proprietor, if resident at home, would consume foreign produce, his absence, not increasing exports, does not affect local industry. So far as the proprietor's absence causes manufactures to be exported, his countrymen are not prejudiced. For they may have as profitable employment in manufacturing those exports as, if the proprietor had resided at home, they would have had in supplying manufactured commodities or services for his use. But if the proprietor by his absence causes raw materials to be exported, while if present he would have used native manufactures and services, his absence tends to deprive his countrymen of employment, to diminish their prosperity, and perhaps their numbers. This reasoning is based on Senior's *Lectures on the Rate of Wages* (Lecture II), and *Political Economy* (pp. 155–61). Senior's position is in a just mean between two extremes – the popular fallacy and the paradox of McCulloch. On the one hand it is asserted that between the payment of a debt to an absentee and a resident there is the same difference as between the payment and non-payment of a tribute to a foreign country.

On the other hand it is denied that there is any difference at all. The grosser form of the vulgar error, the conception that the income of the absentee is drawn from the tributary country in specie, is exemplified in Thomas Prior's *List of Absentees* (1727). McCulloch's arguments are stated in the essay on 'Absenteeism' in his *Treatises and Essays on Money*, etc., and in the evidence given by him before some of the parliamentary commissions which are referred to below. Asked 'Do you see any difference between raw produce and manufactured goods', McCulloch replies, 'I do not think it makes any difference' (compare *Treatises and Essays*, p. 232). He appeals to observation, and finds that the tenants of absentee landlords are 'subjected to less fleecing and extortion than those of residents'.

J.S. Mill attributes to absenteeism a tendency to lower the level of prices in the country from which the absentee draws an income; with the consequence that the inhabitants of that country obtain their imports at an increased cost of effort and sacrifice (*Unsettled Questions*, Essay i, p. 43). Mill's meaning may be made clearer by a study of the rest of the essay which has been cited, and of the parallel passage in his *Political Economy* (Book v, ch. iv, § 6), where he argues that an inequality between exports and imports results in an 'efflux of money' from one country to another.

Upon less distinct grounds Quesnay connects absenteeism with a development of trade and industry in an unhealthy direction (*Oeuvres*, ed. Oncken, p. 189). Among recondite considerations which may bear on the subject should be mentioned Cantillon's theory concerning the effect of the consumption of the rich on the growth of population (*Essai*, pt. i, ch. xv).

II. Other economical advantages lost by absenteeism are those which spring from the interest which a resident is apt to take in the things and persons about him. Thus he may be prompted to invest capital in local improvements, or to act as an employer of workmen. 'It is not the simple amount of the rental being remitted to another country', says Arthur Young, 'but the damp on all sorts of improvements'. D'Argenson in his *Considérations sur le gouvernement ancien et présent de la France* (1765, p. 183), attributes great importance to the master's eye.

The good feeling which is apt to grow up between a resident landlord and his tenantry has material as well as moral results, which are generally beneficial. The absentee is less likely to take account of circumstances (e.g., tenant's improvements), which render rack-renting unjust. He is less likely to make allowance for calamities which render punctual payment difficult. 'Miseries of which he can see nothing, and probably hear as little of, can make no impression' (A. Young). He is glad to get rid of responsibility by dealing with a 'middleman', or intermediate tenant – an additional wheel in the machinery of exaction, calculated to grind relentlessly those placed underneath it. Without the softening influence of personal communication between the owner and the cultivator of the soil, the 'cash nexus' is liable to be strained beyond the limit of human patience, and to burst violently. There can be little doubt but that absenteeism has been one potent cause of the misery and disturbances in Ireland. The same cause has produced like effects in cases widely different in other respects. The cruellest oppressors of the French peasantry before the Revolution were the *fermiers*, who purchased for an annual sum the right to collect the dues of absentee seigneurs. The violence of the Granger Railway legislation in the western states of America is attributed to the fact that the shareholders damnified were absentee proprietors (Seligman, *Journal of Political Science*, 1888).

There are also the moral advantages due to the influence and example of a cultivated upper class. The extent of this benefit will vary according to the character of the proprietors and the people. In some cases it may be, as Adam Smith says, that 'the inhabitants of a large village, after having made considerable progress in manufactures, have become idle in consequence of a great lord having taken up his residence in their neighbourhood'. The opposite view, presented by Miss Edgeworth in her *Absentee*, may be true

in other states of civilization. Perhaps the safest generalization is that made by Senior that 'in general the presence of men of large fortune is morally detrimental, and that of men of moderate fortune morally beneficial, to their immediate neighbourhood'.

Reprinted from *Palgrave's Dictionary of Political Economy.*

References

Brodrick, G.C. 1881. *English land and English landlords*. London.

Carey, H. 1835. *Essay on the rate of wages*. Philadelphia: Carey, Lea & Blanchard.

de Lavergne, L. 1860. *Economie rurale de la France depuis 1789*. Paris: Guillamin.

Levasseur, E. 1885. A summary of the results of the recent Italian Commission. *Journal des Economistes*.

Levasseur, E. 1889. *La population Française.* Paris.

Montchrétien. 1615. *L'économie politique patronale. Traicté de l'oekonomie politique*, ed. Th. Funck-Bretano. Paris, 1889.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. London: W. Strahan & T. Cadell.

Taine, H. 1876. *L'ancien régime.* Paris.

Tocqueville, A. de Clerel. 1856. *L'ancien régime et la révolution*, 3rd edn. Paris, 1857.

Wakefield, E. 1812. *An account of Ireland, statistical and political*. London.

Young, A. 1780. *A tour in Ireland*. London: T. Cannell & J. Dodsley.

Absolute and Exchangeable Value

John Eatwell

Abstract

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

Keywords

Absolute and exchangeable value; Cairnes, F. E.; Cairnes, J. E.; Class; Classical economics; Invariable standard of value; Labour theory of value; Marx, K.H.; Rate of profit; Ricardo, D.; Sraffa, P.; Surplus

JEL Classifications

D0

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

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

Yet the notion of absolute value arose naturally within Ricardo's analysis of value and distribution. The central problem of classical theory is to relate the physical magnitude of surplus (defined as the social output *minus* the replacement of materials used in its production and the wage goods paid to the labourers employed) to the general rate of profit and the rents in terms of

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

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

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

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

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

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

The importance of the essay derives from the reinforcement it provides to that interpretation of Ricardo's theory of value and distribution which suggests that the problem of the determination of the relative values of commodities stemmed from Ricardo's desire to relate his image of the division of social product as a physical magnitude to the wages, rents, and rate of profit of a market economy. Ricardo was not interested for its own sake in the problem of why two commodities produced by the same quantities of labour are not of the same exchangeable value. He was, rather, concerned by the fact that as distribution of social output *changes* exchangeable value *changes,* disrupting and obscuring an otherwise clear vision. It was this emphasis on the fact that *changes* in distribution lead to changes in exchangeable value, even though the quantity of social output and the method by which it is produced are unchanged, which led Ricardo into the intellectual cul-de-sac of the search for an invariable standard of value.

The absolute value of a commodity is the value of that commodity measured in terms of an

invariable standard. An invariable standard of value may be found

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

Changes in the absolute values of commodities could then derive only from changes in the amount of labour embodied in them, and the value of social output would be invariate to its distribution.

Yet precisely because all commodities are not produced under the same circumstances, 'difficulty or facility of production is not absolutely the only cause of variation in value, there is one other, the rise or fall of wages' since commodities cannot 'be produced and brought to market in precisely the same time' (1823, p. 368). Hence Ricardo must conclude, rather sadly, that 'there is no such thing in nature as a perfect measure of value' (1823, p. 404) – there is no such thing as an invariable standard of value.

Marx (1883), who could not, of course, have seen the papers on Absolute and Exchangeable Value, was critical of Ricardo's absorption with the search for an invariable standard. The focus on changes in relative value obscured the fact that commodities do not exchange at rates proportional to their labour values (labour embodied). Yet Marx's attempt to restore clarity to the analysis of distribution by first determining the rate of profit as the ratio of quantities of labour, and then 'transforming' labour values into prices of production, encounters difficulties which derive from exactly the same source as those which bedevilled Ricardo – the difference in production conditions or 'organic composition of capital' of commodities.

The data of classical theory can be used to determine the rate of profit, as Sraffa (1960) has shown. But the determination cannot be 'sequential' – first specifying a theory of value and then evaluating the ratio of surplus to capital advanced by means of that predetermined theory of value. Rather the rate of profit and the rates at which commodities exchange must be determined simultaneously.

See Also

► Ricardo, David (1772–1823)

Bibliography

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

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

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

Sraffa, P. 1951. *Introduction to works and correspondence of David Ricardo*. Vol. 1. Cambridge: Cambridge University Press.

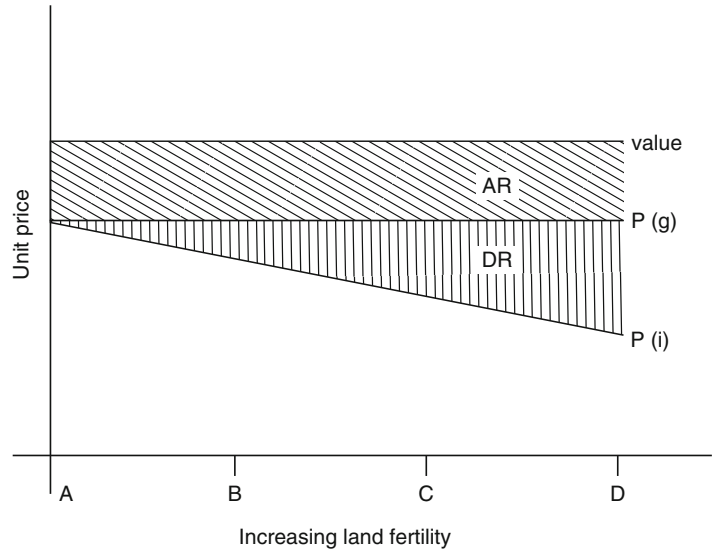
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Absolute Rent

Ednaldo Araquem da Silva

Marx's work on rent was based on his studies of the statistical reports published after the Russian Agrarian Reform of 1861. The importance of the Russian case on Marx's thinking is highlighted in Engels' 'Preface' to the third volume of Marx's Capital, which draws a parallel between the influence of Russia's diverse land tenure system on Marx's analysis of rent and the role of England on his analysis of industrial wage-labour.

Although the economic surplus normally takes the form of profits in the capitalist system, Marx gave considerable attention to rent. In chapter XLV of the third volume of *Capital* (1894), and in his critical comments on Ricardo's theory of rent, published in *Theories of Surplus-Value* (1905), Marx introduced the concept of absolute rent as the rent paid by capitalist tenant farmers to landowners, regardless of the fertility of the rented land.



Absolute Rent, Fig. 1 Marx's concept of absolute rent

Marx (1894, pp. 760, 771; 1905, pp. 244, 392) defined absolute rent as the difference between the value of the agricultural product of the least productive land and the *general* production price, $P(g)$. Absolute rent can absorb the entire [value–$P(g)$] difference or a proportion of this difference. In contrast, differential rent is defined as the difference between the general production price and the *individual* production price, $P(i)$. These concepts are depicted in Fig. 1. By definition, absolute rent is positive even on the worst cultivated land, A, whereas differential rent is zero on A, but then becomes positive and increases with improved land fertility, B, C, and D.

Marx's concept of absolute rent is based on two assumptions: (1) the agricultural organic composition of capital is lower than the average of agriculture and industry; and (2) land is cultivated by capitalist tenant farmers. Assumption (1) implies that the value of an agricultural commodity will be *above* its production price; under assumption (2), landowners will lease land only to those capitalist tenants who can pay absolute rent even on the worst quality and most inconveniently located land.

In contrast to other commodities whose organic composition of capital is lower than the average of agriculture and industry, and thus have their values above their production prices, competition among capitalist producers does not reduce the values of the agricultural products to their production prices. The separation of landowners from tenant operators prevents the equalization of profit rates in agriculture with the single rate prevailing in industry. Landowners are therefore able to seize excess or above average agricultural profits and prevent them from entering the process by which the average profit rate is formed (see Marx 1905, p. 37; Murray 1977).

Under Marx's assumptions, the market price of an agricultural product will include the absolute rent above the general production price.

> If the worst soil cannot be cultivated – although its cultivation would yield the price of production – until it produces something in excess of the price of production, [absolute] rent, then landed property is the creative cause of *this* rise in price (Marx 1894, p. 755).

There has been some confusion as to whether the upper limit of the market price of an agricultural product would be set by its individual value on the worst cultivated land. Marx (1905, p. 332) himself asked: 'If landed property gives the power to sell the product above its [production price], *at* its value, why does it not equally well give the power to sell the product *above* its value, at an arbitrary monopoly price?' Echoing Marx, Bortkiewicz (1911) and, much later, Emmanuel (1972) have also questioned why landlords limit

absolute rent to the excess of value over the production price on the worst cultivated land. They suggest that since landowners have the power to withdraw land from cultivation until the market price covers both the absolute rent and the production price of the highest-cost producer, they could also charge a rent in excess of the corresponding value. In capitalist agriculture, absolute rent has a negative impact because it prevents agricultural prices from falling, and because it removes above average profits, a major source of capitalist technical innovation (see Lenin 1901, pp. 119–29).

Despite some ambiguity in Marx's formulation of absolute rent, his argument is persuasive:

> Although landed property may drive the price of agricultural produce above its price of production, it does not depend on this, but rather on the general state of the market, to what degree market-price exceeds the price of production and approaches the value (Marx, 1894, p. 764, see also p. 762; Murray 1977; Flichman 1977).

According to Marx (1894, pp. 760, 765; 1905, pp. 244, 393), the lower composition of agricultural capital compared to that of industry 'is a *historical* difference and can therefore disappear', and so absolute rent would also tend to disappear as the productivity of agricultural labour approaches that of industry. In this case, the production price of an agricultural product would approach its value and any rent paid by the capitalist tenants would constitute a monopoly rent. The monopoly rent is paid above the value of the agricultural product, and it would thus be limited not by value, as in the case of absolute rent, but by foreign agricultural trade, competition among landowners, and the consumers budget (see Marx 1894, pp. 758, 805, 810; 1905, p. 332).

Marx's theory of absolute rent has been by-passed by the controversy over the transformation of values into production prices, and has been little used as a conceptual device to analyse the effect of landownership on capitalist investment in agriculture or the effect of landownership on agricultural prices. Unfortunately, absolute rent has been neglected by Marxist economists, while it seems to be a favourite *bête noire* among sympathetic critics of Marx, such as Bortkiewicz (1911) and Emmanuel (1972). As a result, absolute rent has an uncertain future as a useful theoretical device, despite the fact that in many countries capitalist agriculture still largely conforms to the two basic assumptions made by Marx more than a hundred years ago.

See Also

- ▶ Land Rent
- ▶ Marx, Karl Heinrich (1818–1883)
- ▶ Rent
- ▶ Unequal Exchange

Bibliography

Bortkiewicz, L. 1911. La teoria della rendita fondiaria di Rodbertus e la dottrina di Marx sulla rendita fondiaria assoluta. In *La Teoria Economica di Marx e altri saggi su Böhm-Bawerk, Walras e Pareto*. Turin: Einaudi, 1971.

Emmanuel, A. 1972. *Unequal exchange*. New York: Monthly Review Press.

Flichman, G. 1977. *La Renta del Suelo y el Desarrollo Agrario Argentino*. Buenos Aires: Siglo Veintiuno Editores.

Lenin, V.I. 1901. The agrarian question and the 'critics of Marx'. In *Collected works*, Vol. V, ed. V.I. Lenin. Moscow: Progress Publishers, 1973.

Marx, K. 1894. *Capital*, Vol. III. Moscow: Progress Publishers, 1971.

———. 1905. *Theories of surplus value*, Part II. Moscow: Progress Publishers, 1968. Murray, R. 1977. Value and the theory of rent: I. *Capital & Class* 1(3): 100–122.

Absorption Approach to the Balance of Payments

David Vines

JEL Classifications
E0

The absorption approach to the balance of payments states that a country's balance of trade will only improve if the country's output of goods and

services increases by more than its absorption, where the term 'absorption' means expenditure by domestic residents on goods and services. This approach was first put forward by Alexander (1952, 1959).

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

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

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

which may be rewritten

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

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

What is needed, and what Alexander helped to provide, is an analysis of exactly how output and absorption change, in response to a devaluation, and indeed in response to other developments in the economy. Such a gap was also being filled at the time by Keynesian writers (Robinson 1937; Harrod 1939; Machlup 1943; Meade 1951; Harberger 1950; Laursen and Metzler 1950; see also Swan 1956).

All of these authors grafted the Keynesian multiplier onto the elasticities approach. The resulting hybrid construct can be used to analyse the effects of a devaluation as follows. Suppose that the price elasticity effects do improve the balance of trade, X–M, by 'switching' expenditures towards domestic goods. Then these 'expenditure-switching' effects provide a positive stimulus to the Keynesian multiplier process, and drive up output Y and absorption $C + I + G$. Let x be the expenditure-switching effects on the trade balance of a devaluation of the currency by one unit, and let the overall effects of this devaluation on the trade balance be y. Let the propensity to consume be c, the tax rate be t and the propensity to import m, so that the Keynesian multiplier is $k = 1/[1 - c(1 - t) + m]$. The increase in output resulting from the devaluation is kx and the increase in absorption is $c(1 - t)kx$. And so

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

If the propensity to consume c is less than unity and the tax rate t is positive then absorption increases by less than output, and, as Eq. (3) shows the trade balance is improved by the devaluation. The above sketch shows how the combination of the elasticities approach and Keynesian theory is able to provide the needed analysis of how output and absorption change following a devaluation. And instead of describing the outcomes in terms of output and absorption, as Alexander did, it is possible to give a more conventional Keynesian description, which would proceed as follows. Since the multiplier

$k = 1/[1 - c(1 - t) + m]$ times the propensity to import m is less than unity, the increase in imports induced by the multiplier, mkx, is less than the positive 'expenditure-switching effects', x, and so the trade balance improves.

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

$$\begin{aligned} z &= k - [1 + (1 - t)ck] \\ &= k - [1 - c(1 - t) + m + c(1 - t)]k \\ &= -mk. \end{aligned} \tag{4}$$

Again this outcome can be described in the more conventional Keynesian way: high government expenditure drives up output by the multiplier, k, and sucks in imports of an amount mk.

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

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

whereas, since output is not to be affected,

$$0 = kx\alpha + k\beta \tag{6}$$

Solving for β from Eq. (6) and substituting into Eq. (5), nothing that $1 - c\,(1 - t) = 1/k - m$, gives

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

Thus the required devaluation is simply $\alpha = w/x$ and substituting in Eq. (6) the required change in government expenditure is simply $\beta = -w$. This states what is obvious: government absorption must be reduced enough to release resource from domestic use – the expenditure-changing component of policy – and the devaluation must ensure that these resources are actually used to improve the trade balance, rather than leading to a fall in domestic output – the expenditure-switching component of policy.

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

Modern balance of payments theory has carried criticisms much further than this. It has shown that the hybrid of the Keynesian multiplier and elasticities approaches is inadequate in providing a full analysis of how output and absorption change. First it does not deal with the inflationary effects of devaluation. But one way in which devaluation depresses absorption relative to output is through engendering rises in costs and prices which depress the real incomes (particularly real wages) of domestic consumers (Diaz Alexandro 1966). Furthermore, devaluation may also engender a wage-price spiral so strong as to preserve the real incomes of domestic consumers, with the end result that prices rise by the full extent of the devaluation and there is no relative price change for the price elasticities effects to work on (Ball et al. 1977). In that case positive effects of devaluation on the trade balance can *only* emerge as a result of the effects of higher

prices on absorption (Higher prices lower the real wealth of consumers and perhaps also increase the tax burden if tax rates are progressive and not indexed with inflation). Second, the multiplier-plus-elasticities analysis is not appropriate in analysing the effects of a devaluation not accompanied by any expenditure changing policy if the economy is at full employment, for in that case output cannot be expanded through the multiplier, and the effects of the devaluation must primarily work through the influence of inflation on absorption described above. Third, the multiplier-plus-elasticities analysis does not deal with monetary conditions. A devaluation, because it raises prices, may initially also cause higher interest rates which helps to curtail absorption. But if the improvement in the trade balance caused by the devaluation is allowed to lead to an expansion of the domestic money supply, then gradually interest rates will fall, absorption will rise, and the effects of the devaluation may turn out to be temporary. This issue has been analysed by the Monetary Approach to the Balance of Payments (Frenkel and Johnson 1976; Kyle 1976; McCallum and Vines 1981). Alexander made many of these points in his articles whereas the authors cited at the end of the fourth paragraph tended to skate over them. For that reason his work prefigures much subsequent balance of payments theory.

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

See Also

- ► Elasticities Approach to the Balance of Payments
- ► Monetary Approach to the Balance of Payments

Bibliography

Alexander, S.S. 1952. Effects of devaluation on a trade balance. *International Monetary Fund Staff Papers* 2: 263–278.

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

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

Caves, R.E., and H.G. Johnson, eds. 1968. *Reading in international economics*. London: George Allen & Unwin.

Diaz Alexandro, C. 1966. *Exchange rate devaluation in a semi-industrialized country: The experience of Argentina 1955–1961*. Boston: MIT Press.

Frenkel, J.A., and H.G. Johnson, eds. 1976. *The monetary approach to the balance of payments*. London: George Allen & Unwin.

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

Harrod, R.F. 1939. *International economics*, Cambridge economic handbooks, VIII. 2nd ed. London: Nisbet & Co..

Kyle, J.F. 1976. *The balance of payments in a monetary economy*. Princeton: Princeton University Press.

Laursen, S., and L. Metzler. 1950. Flexible exchange rates and the theory of employment. *Review of Economics and Statistics* 32: 281–299.

Machlup, F. 1943. *International trade and the national income multiplier*. Philadelphia: Blakiston Co.

McCallum, J., and D. Vines. 1981. Cambridge and Chicago on the balance of payments. *Economic Journal* 91: 439–453.

Meade, J.E. 1951. *The balance of payments*. London: Oxford University Press.

Robinson, J.V. 1937. The foreign exchanges. In *Essays in the theory of employment*, ed. J. Robinson, 2nd ed, 1947. Reprinted in *Reading in the theory of international trade*, ed. H.S. Ellis and L.A. Metzler, 83–103. Philadelphia: Blakiston, 1949.

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

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

Absorptive Capacity

Richard S. Eckaus

The idea that the productivity of new investment is a declining function of the rate of investment – the

concept labelled 'absorptive capacity' – has attracted attention in development economics because of its implications as a constraint on growth.

The hypothesis began to emerge most clearly in the 1950s in the form of a limit on the total amount of investment which could be carried out and/or used in any period, as if the marginal productivity of resources devoted to investment would, at some level of total investment undertaken, fall to zero. This was the position taken by Horvath (1958), citing experience in Yugoslavia and Eastern Europe. An Economic Commission for Asia and the Far East (ECAFE) report claimed that 'capacity sets a limit to the amount of efficient investment physically possible', introducing the distinction between 'efficient' and, presumably, 'inefficient' investment (ECAFE 1960). In the early discussions, the concept was used to represent all the constraints on development which economists could not easily put into the conventional production function, 'the supply of skilled labour, administrative capacity, entrepreneurship and social change' (Marris 1970).

Rosenstein-Rodan (1961), Adler (1965) and others described the absorptive capacity concept as a relationship between the productivity and the rate of investment, rather than as an absolute ceiling on investment's productivity. The source of the relationship were not discussed in depth nor investigated empirically and it remained a 'black box' whose inner workings were never fully explained. Nonetheless, by the mid-1960s the absorptive capacity idea had become a part of the standard toolbox of development economics and used readily to explain difficulties experienced in attempts to accelerate economic growth.

Research on growth and planning models led to both a refinement of the concept and new speculation about its sources. Kendrick and Taylor (1969), following a suggestion by Dorfman and Thoreson (1969), modelled the absorptive capacity constraint as a permanent reduction in the productivity of new investment related to the rate of investment, as if an increase in investment were accompanied by the use of progressively inferior engineering design and materials. Eckaus (1972) formulated the constraint by making the productivity of successive tranches of investment in any year decline relative to the original tranche with, however, the decline only being temporary. In subsequent periods after the new capital was completed, its productivity would grow to 'rated' levels. He offered the hypothesis that, as investment increases, less and less well qualified engineers and workers and less suitable equipment are employed in producing the new capital goods and bringing them into production.

The absorptive capacity concepts came to play a critical role in the economy-wide policy models which were formulated as linear programming problems. If the objective function in such models is linear, for example, the simple discounted sum of aggregate consumption over the plan period and, if all the constraints are linear and do not control the timing of consumption, the solutions of the models will exhibit 'flip-flop' or 'bang-bang' behaviour. Aggregate consumption will be concentrated either at the beginning or at the end of the planning period. This unrealistic and undesirable result can be controlled by constraints on the timing of consumption (Eckaus and Parikh 1968). An aggregate utility function with declining marginal utility as a nonlinear objective function and/or absorptive capacity constraints, which are essentially nonlinear relations between investment and increments in output, are, however, theoretically more satisfactory means of avoiding 'bang-bang'.

The absorptive capacity concept is related closely to a generalization which emerged quite independently of the development literature from the study of factors constraining the growth of firms in advanced countries (Penrose 1959). This was embodied in a theoretical growth model by Uzawa (1969). The concept is also a close relation, if not the twin, of an idea which appeared early in the macroeconomic analysis literature only to be lost and then revived once more. In chapter 11 of the *General Theory*, Keynes describes the marginal efficiency of capital, that is, the productivity of new investment, as declining with the rate of new investment because, 'pressure on the facilities for producing that type of capital will cause its supply price to increase' (Keynes 1936). Under the title of 'adjustment costs', this characterization began to figure

prominently in the macroeconomic literature in the late 1960s (Lucas 1967).

'Adjustment costs' is a phrase which is as appealing as 'absorptive capacity'. The phenomenon is not explained by giving it a name, however. While the fact that economists continue to resort to the idea might be counted as evidence that it reflects a reality, the empirical research on its sources is still limited.

See Also

- ▶ Adjustment Costs
- ▶ Development Economics

References

Adler, J. 1965. *Absorptive capacity and its determinants*. Washington, DC: Brookings Institution.

Dorfman, R., and R. Thoreson. 1969. *Optimal patterns of growth and aid with diminishing returns to investment and consumption*, Economic Development Report, vol. 142. Cambridge, MA: Development Research Group, Harvard University.

Eckaus, R.S. 1972. Absorptive capacity as a constraint due to maturation processes. In *Development and planning: Essays in honour of Paul Rosenstein-Rodan*,-ed. J. Bhagwati and R.S. Eckaus. Cambridge, MA: MIT Press.

Eckaus, R.S., and K.S. Parikh. 1968. *Planning for growth*. Cambridge, MA: MIT Press.

Economic Commission for Asia and the Far East (ECAFE). 1960. *Programming techniques for economic development*. Bangkok: United Nations.

Horvath, B. 1958. The optimum rate of investment. *Economic Journal* 68: 747–767.

Kendrick, D.A., and L.J. Taylor. 1969. A dynamic nonlinear planning model for Korea. In *Practical approaches to development planning*, ed. I. Adelman. Baltimore: Johns Hopkins Press.

Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.

Lucas, R. 1967. Adjustment costs and the theory of supply. *Journal of Political Economy* 75: 321–334.

Marris, R. 1970. Can we measure the need for development assistance? *Economic Journal* 80: 650–668.

Penrose, E. 1959. *The theory of the growth of the firm*. Oxford: Blackwell.

Rosenstein-Rodan, P.N. 1961. International aid for underdeveloped countries. *Review of Economics and Statistics* 43(2): 107–138.

Uzawa, H. 1969. Time preference and the Penrose effect in a two-class model of economic growth. *Journal of Political Economy* 77: 628–652.

Abstinence

N. De Marchi

'Abstinence' was Nassau Senior's term for that conduct for which profit is the reward (1836, p. 59). He meant it to convey two things: 'both the act of abstaining from the unproductive use of capital, and also the similar conduct of the man who devotes his labour to the production of remote rather than immediate results' (1836, p. 89). The term he knew was not ideal, but it was preferable to 'providence', which implies nothing of self-denial; and to 'frugality', which implies care and attention, that is, labour, which analytically Senior wanted to keep distinct from the agent of production rewarded by profit. For the same reason he chose not to speak of profit in relation to 'capital'. Capital usually combines the services of natural agents, labour and abstinence, but it is desirable in an analysis to keep their several contributions distinct.

Despite the desirability of precision in analysis, Senior had to admit that in practice 'it is often difficult to distinguish profit from wages', or, for that matter, rent from profit (1848–9, pp. 149–50). Nor was he, nor any of the other classical writers who took over his terminology (e.g. J.S. Mill 1848, p. 34), able to quantify the reward of abstinence. Clearly it is the minimum return for there to be any accumulation, and, since profit is an uncertain expectation (1836, p. 187), must be at least equal to the rate of interest on a government bond; but beyond that exactly how the rate is settled was not paid much attention. It must not be thought, however, that profit is just the reward for the initial refraining from consuming one's capital. That would make any net return after the first period simply rent. The fact that Senior stressed abstinence also in relation to activity with remote results, suggests that he was fully aware that profit must be calculated as the present value of a stream of returns.

The notion of abstinence has been regarded by Marxian writers as a poor apology for a

justification of the payment of interest. They have ridiculed it, using examples comparing the 'abstinence' of a Rothschild with the profligacy of a labourer who spends all his meagre income. Even John Stuart Mill, in his draft *Chapters on Socialism* wrote: 'The very idea of distributive justice, or proportionality between success and merit, or between success and exertion, is in the present state of society so manifestly chimerical as to be relegated to the regions of romance' (1879, p. 714).

These sentiments are misleading in relation to Senior's deployment of 'abstinence'. The idea derives from his Stoic perspective on supply. Production involves overcoming obstacles such as a natural preference for leisure and for present enjoyment; hence prudent behaviour to counter these impediments requires and merits recompense. Abstinence, for Senior, was on a par with the other agents of production – labour and natural agents – and is critical to one of his four fundamental propositions of economic science, the notion that the powers of labour 'may be indefinitely increased by using their Products as the means of further Production' (1836, pp. 26, 58). Senior's point is simply that abstinence is a necessary precondition for capital to emerge.

Marshall, with characteristic appositeness, insisted on a distinction between abstemiousness and waiting, and used the latter to replace abstinence (1890, pp. 232–3). He also saw the important point in Senior's discussion, namely, the need to encourage the 'faculty of realizing the future' or 'man's *prospectiveness*' (ibid., p. 233). Without encouragement to this faculty there will be no supply of capital. Others, such as Böhm-Bawerk and Fisher, argued against treating abstinence as a cost (Fisher 1907, pp. 43–5). But this criticism scarcely touches Senior, and Fisher is basically at one with Marshall in stressing prospectiveness. Fisher's emphasis, however, is on time-preference and the fact that time-preference itself will depend on the size, distribution over time, composition and probability of the prospective income stream facing an individual (ibid., pp. 92–4). This is the natural link with recent models embodying inter-generational transfers and infinite time-horizons.

See Also

- ▶ Senior, Nassau William (1790–1864)
- ▶ Waiting

Bibliography

Fisher, I. 1907. *The rate of interest*. New York: Macmillan.

Marshall, A. 1890. *Principles of economics*, vol. 2, 9 (variorum)th ed. London: Macmillan for the Royal Economic Society, 1961.

Mill, J.S. 1848. Principles of political economy. In *Collected works of John Stuart Mill*, vol. II, III, ed. J.M. Robson. Toronto: University of Toronto Press, 1965.

Mill, J.S. 1879. Chapters on socialism. In *Collected works*, Vol. V, 1967.

Senior, N.W. 1836. *An outline of the science of political economy*, Reprint. New York: Kelley, 1965.

Senior, N.W. 1848–9. *Course of lectures delivered in the University of Oxford*. Unpublished; quoted in *Industrial efficiency and social economy* by Nassau W. Senior, ed. S. Leon Levy, 2. vols, New York: Henry Holt & Co., 1928.

Abstract and Concrete Labour

Anwar Shaikh

The reproduction of society requires the production and distribution of the mass of products which forms the material basis of its existence. This in turn means that each society must somehow ensure that its available social labour time is regularly directed, in particular quantities and proportions, towards the specific applications needed to ensure social reproduction. As Marx points out, 'every child knows that a nation which ceased to work … even for a few weeks, would perish' (Marx 1867a).

The above implies that all labour has two distinct aspects. As a part of the general pool of society's labour, it is merely one portion of the human energy available to the community. In this respect all labour is essentially the same, representing the expenditure of 'human labour-power in general' in its capacity as simply one

part of the division of general social labour. This is labour as *social labour*. But at the same time, individual labour occurs in the form of a specific activity aimed at a specific result. Here it is the particular quality of the labour, its determination as labour of mining, metalworking, weaving, distribution, etc., which is relevant. This is labour as *concrete labour*, related to the concrete result of its activity.

Although the dialectic between concrete and social labour is a necessary part of social reproduction, their inter-connection is hard to discern within societies which produce things-for-exchange (commodities), because in this case individual activities are undertaken without any apparent consideration for the necessity of a social division of labour. All useful objects now appear to be naturally endowed with quantitative worth in exchange (*exchange value*), and this apparently natural property in turn seems to regulate the actual division of labour.

It is at this point that Marx introduces two crucial questions. What precisely is a commodity? And more importantly, why does it become socially necessary to attach an exchange value to it? He begins his answer by observing that as a useful good a commodity is simply a concrete bundle of different socially desirable properties. In this respect it is similar to particular, qualitatively distinct useful objects in all social forms of organization. But as an exchangeable good, its salient property is that it is treated socially as being qualitatively *identical* to every other commodity. This is manifested in the fact that when commodities are assigned differing quantities of exchange value, expressed in some common measure, they are thereby being socially regarded as qualitatively alike, all reducible to the same homogenous measure of quantitative worth. A commodity is therefore a doublet of opposite characteristics: a multiplicity of concrete useful properties (use value) on the one hand, and a single magnitude of homogenous quantitative worth (exchange value) on the other.

The double character of a commodity is strikingly reminiscent of the previously noted duality of labour as particular concrete labour and as general social labour. Indeed, in commodity producing society the various concrete labours 'only count as homogeneous labour when under *objectified husk*', that is, when they 'relate to one another as human labour by relating *their products to one another as values*'. The concrete labours are thus counted as social labour only when they are *valorized*, and the necessity of exchange value lies precisely in the fact that it is through this device that a society containing apparently independent private producers comes to grips with the social content of their individual labours. To answer Marx's second question, exchange value is the particular historical mode of expressing the general necessity of social labour.

The notion that exchange value is a historically specific way of accounting for social labour time does not imply that the terms of exchange of commodities always reflect the quantities of valorized social labour time that went into their respective production. Indeed, Marx distinguishes between the case in which particular useful objects are produced for direct use and only accidentally or occasionally find their way into the sphere of exchange, and the case in which goods are produced *in order* to be exchanged. In the first case, when for example otherwise self-sufficient tribes occasionally barter a few of their products, the relation between concrete labour and social labour is effectively determined within each social group, and exchange merely serves to create a temporary equivalence between the respective social labours involved. Because the objects in question are produced as useful objects and become commodities only when they enter exchange, the labours involved are valorized only in exchange itself. Moreover, since these activities do not depend fundamentally on exchange (and hence on the valorization of their labour), the precise conditions of exchange can in turn be decided by a variety of factors, ranging from broad structural influences to merely conjunctural or even accidental ones.

At the opposite extreme is the case of goods produced solely for exchange. Now, the particular labours involved are *aimed* at producing exchangeable goods, and the valorization of

these labours is an intrinsic part of their reproduction. As producers of commodities, these labours create not only bundles of useful properties (use-values), but also amounts of abstract quantitative worth. In the former aspect, they are of course concrete labours; but in the latter, they are *value creating* activities whose content as social labour is manifest only in-and-through the abstract quantitative worth of their products. To emphasize this particular historical form of the duality of labour, Marx identifies that labour which is engaged in the production of commodities as being both concrete (use-value creating) labour, and *abstract* (value creating) labour.

Three further points must be briefly mentioned. First of all, Marx argues that abstract labour time not only stands behind the production of commodities, but that the magnitudes of these labour times actually regulate the exchange relations of these commodities. To this end, he defines the quantity of abstract labour 'socially necessary . . . to produce an article under the normal conditions of production' as the (inner) *value* of the commodity, since it is the 'intrinsic measure' of the exchange value. Secondly, he distinguishes between the conditions under which the exchange relations of commodities are dependent on their (labour) values, and the conditions in which they are controlled by them. It is only in the latter instance, in which capitalism has effectively generalized commodity production, that the reproduction of society is regulated by the law of value. Lastly, he notes that once commodity production is indeed generalized, so that social labour appears only under objective husk, then the social relation among producers is actually regulated by the mysterious value-relation between their products. In this topsy turvy world, a social relation among persons appears in their eyes to be in fact a relation among things. This is what Marx calls the Fetishism of Commodities which is characteristic of capitalism.

See Also

- ▶ Adaptive Expectations
- ▶ Adjustment Costs
- ▶ Labour Power
- ▶ Marxist Economics
- ▶ Value and Price

References

Marx, K. 1867a. *Capital*, vol. I, 1st ed., ch. 1 and Appendix to ch. 1. In *Value: Studies by Karl Marx.* Trans. and Ed. A. Dragstedt. London: New Park Publications, 1976.

Marx, K. 1867b. *Capital*, vol. I. Introduced by E. Mandel. London: Penguin, 1976, ch. 1.

Marx, K. 1879. Marginal notes to A. Wagner's *textbook on political economy*. In *Value: Studies by Karl Marx*. Trans. and Ed. A. Dragstedt. London: New Park Publications, 1976.

A

Acceleration Principle

P. N. Junankar

Abstract

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

Keywords

Acceleration principle; Aftalion, A.; Aggregate demand; Aggregate investment; Business cycles; Capital–output coefficient; Chenery, H. B.; Clark, J. M.; Depreciation; Derived demand; Distributed lag accelerator; Eisner, R.; Expectations; Haberler, G.; Harrod, R. F.;

Harrod–Domar growth model; Marx, K. H.; Pigou, A.C.; Technical change

JEL Classifications
E22

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

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

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

$$K_t^* = vY_t$$

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

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

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

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

In other words, for net investment to be positive, output must be growing: v is called the accelerator.

The acceleration principle can be derived from a cost-minimizing model on the assumption of either fixed (technical) coefficients and exogenous output, or variable coefficients with constant relative prices of inputs and exogenous output.

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

important qualification to the simple accelerator model was than an increase in (expected) output would lead to an increase in investment only if it was believed that, in some way, the increase was 'permanent' or at least of long duration.

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

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

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

In evaluating the acceleration principle it is worth stressing that, in some versions, it is used as an explanation of investment demand with the implicit assumption that the supply of capital goods will always satisfy that demand. In models where the acceleration principle is used to explain the *supply* of capital goods, it is assumed that they always satisfy the demand for them. The flexible accelerator is a hybrid version which includes both demand and supply elements. Although there is no formal treatment of replacement investment, it is usually postulated to be determined in the same way as net investment. A major shortcoming of the acceleration principle is its simplistic treatment of *expectations* of future demand as well as its neglect of expectations of the time paths of output and input prices. Although most of the work in this field treats the acceleration principle as applying to the aggregate economy, it has also been used to explain investment by firms. It is especially important that the supply of capital goods is formally modelled along with the acceleration principle determining investment demand. Aggregation over firms is usually assumed to be a simple exercise of 'blowing up' an individual firm's investment demand. However, it should not be forgotten that in a modern capitalist economy an individual firm may invest by simply taking over an existing firm rather than by buying new capital goods. An important shortcoming of the acceleration principle is its neglect of technological change.

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

See Also

- ▶ Clark, John Maurice (1884–1963)
- ▶ Multiplier–Accelerator Interaction

Bibliography

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

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

Clark, J.M. 1917. Business acceleration and the law of demand: A technical factor in economic cycles. *Journal of Political Economy* 25: 217–235.

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

Eisner, R., and R. Strotz. 1963. Determinants of business investment. In *Commission on money and credit, Impacts of monetary policy.* Englewood Cliffs: Prentice-Hall.

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

Harrod, R.F. 1936. *The trade cycle*. Oxford: Oxford University Press.

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

Knox, A.D. 1952. The acceleration principle and the theory of investment: A survey. *Economica* 19 (75): 269–297.
Koyck, L. 1954. *Distributed lags and investment analysis*. Amsterdam: North-Holland.
Marx, K.H. 1863. *Theories of surplus value*, Part II. Moscow: Progress Publishers.
Pigou, A.C. 1927. *Industrial fluctuations*. 2nd ed. London: Macmillan, 1929.

Access to Land and Development

Alain de Janvry and Elisabeth Sadoulet

Abstract

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

Keywords

Access to land and development; Agriculture and economic development; Common property resources; Democracy and economic development; Family farm system; Geographical information systems; Human capital; Inheritance rights; Land markets; Land reform; Land registration; Land use; Natural capital; Open-access resources; Poverty alleviation programmes; Property rights; Rent control; Rental markets; Total factor productivity; Tragedy of the commons

JEL Classifications

O1

Access to land, and the conditions under which it happens, play a fundamental role in economic development. This is because the way the modes of access to land and the rules and conditions of access are set, as policy instruments, has the potential of increasing agricultural output and aggregate income growth, helping reduce poverty and inequality, improving environmental sustainability, and providing the basis for effective governance and securing peace. This potential role is, however, difficult to capture, and there are many cases of failure. History is indeed replete with serious conflicts over access to land and with instances of wasteful use of the land, both privately and socially. Governments and development agencies have for this reason had to deal with the 'land question' as an important item on their agendas (de Janvry et al. 2002). We explain in this article: (a) why access to land, and the conditions under which it is accessed and used, are important for economic development, (b) how different types of property rights can affect access and use, (*c*) the different modes of access, and in particular the role of land markets, and (*d*) some of the policy implications, in order to show how access to and use of the land can contribute to economic development. We stress in this article that access to land may be a difficult policy question, but that access will translate into development only if the harder question of influencing the way it is used is effectively resolved.

Importance of Access to Land for Development

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

If land is to serve as an instrument for output and income growth, investments have to be made to improve its productivity. For this to happen, incentives have to be provided. Some of these

investments are short-term, but many others are tied to the land for long periods of time. As a result, security of access is a central policy issue as it is necessary for these investments to be made. Security can be guaranteed through formal means such as titles and legal enforcement, but also through informal mechanisms such as community recognition and enforcement of rights. Whichever way it is achieved, security of access must be credible if it is to induce investment (Deininger 2003).

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

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

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

Property Rights Over Land

The benefits that can be derived from access to land depend on the property rights that codify access and use. Property rights become increasingly complete as they allow the following functions to accumulate: entry, extraction, management, exclusion, and sale (Ostrom 2002). Open-access resources grant to all the rights of entry and extraction. They typically induce overextraction, leading to the 'tragedy of the commons'. Common property resources grant to members of a defined group, such as a community, the rights of entry, extraction, management, and exclusion of non-community members. This form of property right can result in socially optimal resource use if community members have the ability to cooperate in defining and enforcing rules for individual extraction and maintenance (Baland and Platteau 1996). Public ownership with centralized management also gives leaders these same rights. Socially optimum resource use can be achieved if controls and incentives can be aligned between leaders and workers, which has historically proved to be difficult in agriculture, despite many attempts. Finally, individual or corporate property rights give owners the full bundle

of rights, including those of rental and sale. The effectiveness of this form of property right in land use depends on the existence of efficient land rental and sales markets, as well as the ability to internalize externalities, achieve economies of scale, and access mechanisms for risk spreading. Common property resources with cooperation may be a superior form of property right when individual tenures are unable to fulfil these functions.

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

Modes of Access to Land

With open-access resources, entry is granted to all. Access to common property resources is usually given by birthright in a particular community. Clear demarcation of boundaries and clear determination of membership are important to permit the definition and enforcement of rules. Individual encroachment on public lands and establishing adverse possession rights through occupation is an important form of access where public lands remain plentiful. Finally, individual inheritance is also one of the most prevalent forms of access to land, with eventually discriminatory rights due to primogeniture and to gender and kinship privileges in inheritance.

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

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

Access to Land and Development: Policy Implications

In managing their ‘land question’, most countries have experimented with some type of land reform programme (Dorner 1992). This includes land reforms that have used the threat of expropriation to induce extensively used large farms to modernize or subdivide into smaller farms (Brazil). Other reforms have collectivized the land, either as state farms or as cooperatives. This has generally, as in Russia and eastern Europe, been based on the belief in economies of scale in farming and the

superior efficiency of centralized management. In other cases, as in Latin America, collective farms have been used to facilitate transitions between large haciendas and subsequent distribution of the land as individual tenures (Mexico, Peru, Chile). Finally, the inverse relation between total factor productivity and farm size has been invoked in implementing redistributive land reforms that have established family farms out of former large farms (Taiwan, South Korea) or out of state farms or cooperatives (Albania, Bulgaria).

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

Conclusion

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

See Also

- ▶ Common Property Resources
- ▶ Land Markets
- ▶ Peasant Economy
- ▶ Poverty Alleviation Programmes
- ▶ Property Rights
- ▶ Tragedy of the Commons

Bibliography

Agarwal, B. 1994. *A field of one's own: Gender and land rights in South Asia*. Cambridge, MA.: Cambridge University Press.

Baland, J.-M., and J.-P. Platteau. 1996. *Halting degradation of natural resources: Is there a role for rural communities?* Oxford: Oxford University Press.

Bardhan, P. 1984. *Land, labor, and rural poverty: Essays in development economics*. New York: Columbia University Press.

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

de Janvry, A., G. Gordillo, J.-P. Platteau, and E. Sadoulet. 2002. *Access to land, rural poverty, and public action*. Oxford, UK: Oxford University Press.

Deininger, K. 2003. *Land policies for growth and poverty reduction*. Washington, DC: World Bank.

Dorner, P. 1992. *Latin american land reform in theory and practice: A retrospective analysis*. Madison: University of Wisconsin Press.

Hayami, Y., and K. Otsuka. 1993. *The economics of contract choice: An agrarian perspective*. Oxford, UK: Oxford University Press.

Ostrom, E. 2002. The puzzle of counterproductive property rights reforms: A conceptual analysis. In *Access to land, rural poverty, and public action*, ed. A. de Janvry et al. Oxford, UK: Oxford University Press.

Warriner, D. 1969. *Land reform in principle and practice*. Oxford: Clarendon Press.

Accounting and Economics

Joel S. Demski

Abstract

Accounting provides an important source of economic measures, yet consistently falls

short of the economist's conceptual ideal. This shortfall is fodder for economic research, is the result of economic forces, and is the key to making the best possible use of these measures.

Keywords

Accounting and economics; Auditing regulation; Depreciation; Financial Accounting Standards Board (FASB); Generally Accepted Accounting Principles (GAAP); Historical-cost accounting; Information school of accounting; International Accounting Standards Board (IASB); Measurement school of accounting

JEL Classifications

M4

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

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

This disparity is related to the institutional setting in which accounting products are produced, and to the economic forces operating on and within those institutions.

Institutional Highlights

Accounting cannot be divorced from its institutional setting. Were firms truly single-product entities, and were markets complete and perfect, economic measurement would be well defined, the nirvana of classical income measurement (for example, Hicks 1946) would be operational. Unfortunately, in such a setting no one would pay for the services of an accountant simply because the underlying fundamentals would be assumed to be common knowledge. But firms are multi-product entities, markets are neither perfect nor complete, and the underlying fundamentals are far from common knowledge. Here we find a demand for accounting services, such as measuring a firm's periodic income, the performance of the divisions within that firm, and the cost of each of its products. We also find considerable ambiguity over how best to perform those services.

Firms' published financial reports are the most visible accounting product. They entail a reporting entity (the organization about which the financial reports purport to speak), a listing of resources and obligations in its balance sheet, and a listing of the flow of resources during the reporting period in its income statement. Ambiguity is omnipresent. The reporting entity is not an economically defined firm, as its economic relationships are likely to be more extensive than those identified by its formal reporting; for example, implicit economic arrangements are generally ignored in these reports. Nor is the reporting entity simply a legally defined firm, as it often includes, say, a number of wholly or partially owned though legally free-standing legal entities aggregated into its public reports. Even with an unambiguous reporting entity, that entity's control of economic

resources would be incompletely and inaccurately measured. Some assets, such as proprietary knowledge or capital assets acquired through lease arrangements, would not be included. And among those included we would find a mixture of current prices (for example, cash and some financial instruments) and historical cost (for example, most real assets).

The flow measure is equally ambiguous. It is broadly based on what customers have paid minus the resources that were consumed in the process of satisfying those customers. Such wide-ranging phenomena as product warranties and potential product liabilities, uncollectible accounts, pension plans, advertising, research and development and employee training render precise identification of what customers have paid or what resources were consumed largely the product of art as opposed to science.

Regulation, to no one's surprise, now enters the picture. Public financial reports are typically required to be produced according to Generally Accepted Accounting Principles (GAAP). These reports are also typically required to be audited, where the auditor attests to the claim the reports are in compliance with GAAP. One reason for regulations is that the noted ambiguity places a premium on coordinated measurement approaches, a classic example of a network externality (Wilson 1983). A second reason, based on investor protection concerns and again related to the ambiguity, is the potential for opportunism. Absent auditing, the public financial report is simply management's self-report of its financial results and the unverified claim that those results were measured according to GAAP. Of course the auditor's verification is statistical and judgemental; to no one's surprise, the auditor himself is also regulated.

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

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

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

The Measurement School

The measurement school takes its cue from classical economics. In a fully developed general equilibrium model, with complete and perfect markets (for example, Debreu 1959), value and income are well defined, as is the value of a firm's assets and obligations. The measurement school takes this as a desideratum and emphasizes the importance of approaching this economic ideal reasonably well.

This is the source of accounting's intellectual history, its underlying definitions of asset, liability, income, revenue and expense, and the rhetoric used by its regulators. (Important contributors to this school of thought include Paton 1922; Clark 1923; Canning 1929; Edwards and Bell 1961; Solomons 1965; Chambers 1966).

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

Likewise, with the advent of financial engineering it is natural, from the measurement school perspective, that GAAP require fair value (that is, as if market value) estimates of these instruments. In short, with the measurement school we at least know what it is, conceptually, we are trying to measure.

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

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

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

The Information School

The information school, in contrast, focuses on these economic forces and takes its cue from the economics of uncertainty. It views the accounting product not literally as measures of resources but as information that purports to inform about these resources. Abstractly, then, accounting is a mapping from underlying acts and events into the real numbers. In this view, accounting is one among many sources of information. Analysts, the financial press and trade associations are familiar sources of financial information, as are government statistics themselves. Moreover, firms often engage in voluntary disclosures; for example, new product announcements, major investment announcements, and even so-called earnings warnings where they reveal that a forthcoming earnings measure will be lower than originally anticipated. In addition, the typical financial report reports cash flow, an utterly reliable, unambiguous measure. (Important contributors to this school of thought include Butterworth 1972; Feltham 1972; Ijiri 1975; Beaver 1998; Christensen and Demski 2002).

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

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

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

Conclusion

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

See Also

- ▶ Assets and Liabilities
- ▶ Capital Measurement
- ▶ Cost Functions
- ▶ Depreciation
- ▶ Double-Entry Bookkeeping

- Human Capital
- Measurement
- Pensions
- Present Value

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

Bibliography

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

Ball, R., and P. Brown. 1968. An empirical evaluation of accounting income numbers. *Journal of Accounting Research* 6: 159–178.

Beaver, W. 1998. *Financial reporting: An accounting revolution*. Englewood Cliffs: Prentice-Hall.

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

Canning, J. 1929. *The economics of accountancy*. New York: Ronald Press.

Chambers, R. 1966. *Accounting, evaluation and economic behavior*. Englewood Cliffs: Prentice-Hall.

Christensen, J., and J. Demski. 2002. *Accounting theory: An information content perspective*. New York: McGraw-Hill/Irwin.

Clark, J. 1923. *Studies in the economics of overhead costs*. Chicago: University of Chicago Press.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*. New Haven: Yale University Press.

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

Edwards, E., and P. Bell. 1961. *The theory and measurement of business income*. Berkeley: University of California Press.

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

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

Ijiri, Y. 1975. *Theory of accounting measurement*. Sarasota: American Accounting Association.

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

Mock, T. 1976. *Measurement and accounting information criteria*. Sarasota: American Accounting Association.

Morgenstern, O. 1965. *On the accuracy of economic observations*. Princeton: Princeton University Press.

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

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

Ross, S. 2004. *Neoclassical finance*. Princeton: Princeton University Press.

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

Wilson, R. 1983. Auditing: Perspectives from multiperson decision theory. *The Accounting Review* 58: 305–318.

Accumulation of Capital

Edward J. Nell

The accumulation of capital has been analysed by economists in two very different ways. The most common has been to see it as the expansion of the productive potential of an economy with a given technology, which may be improved in the process. But it has also been understood as the outright transformation of the technical and productive organization of the economy. The first approach leads to analyses based on the idea of steady growth, subsuming the concerns of the second under the heading of 'technical progress'. Such an approach rests on a conception of capital as productive goods or, in more sophisticated versions, as a fund providing command over productive goods. This is not wrong; it is merely inadequate. Capital must also be understood as a way of organizing production and economic activity, so that the accumulation of capital is the extension of this form of organization into areas in which production, exchange and distribution were governed by other rules. This conception of capital emphasizes the importance of organization; so understood, technology and engineering are not abstract science, they are ways of organizing production, and so have an institutional dimension. Accumulation then implies the transformation of institutions as well as production, and steady growth is not applicable (except perhaps as a benchmark).

Besides the distinction between steady state and transformational growth, there is another

principal division in the way that economists have thought about accumulation. One side sees it as 'ploughing back' part of the surplus arising from production; the other as the process of adjusting a scarce resource to its optimal uses, as determined by the market. According to the classical 'surplus' approach, accumulation consists of the productive investment of part of society's net product – the surplus of output over necessary consumption and the requirements for maintaining capital intact – in order to expand productive capacity to take advantage of new or developing markets. The study of accumulation, therefore, needs to explain both the availability of the surplus and the motivation for ploughing it back, and this can be examined either as steady state expansion or as part of a process of transformation.

The originators of the classical tradition saw accumulation as a transformation of the economy. Smith stressed institutional changes, in particular the development of markets and the removal of state barriers, but his analytics were incomplete and partially incorrect. Ricardo offered only a rudimentary explanation of the surplus, in the 'iron law of wages'; accumulation, however, he saw as the natural activity of capitalists, although it would be limited by the rise of food prices caused by the extension of cultivation to marginal lands, shifting distribution in favour of rent. Marx located the origin of the surplus in the exploitation of labour and found the cause of the tendency of the rate of profit to fall in the interaction of competition and technological advance rather than in pressure on marginal land. Each offered a picture with a grand sweep, painted in large strokes. Modern 'surplus' theory is more circumspect and less interesting.

In most modern work accumulation is studied in the context of steady growth. Growth can be aimed at a specific target, or can continue indefinitely. The first is the subject of 'turnpike' studies (so called because to reach a target set of outputs most rapidly the economy first shifts to the balanced growth path – the 'turnpike' – and speeds along it, changing to the desired production mix when it reaches the right size), while the latter is analysed by models in which equilibrium paths of perpetual expansion are determined and their properties examined. So, given a system of production, we ask how that system can be set up so as to grow either over the indefinite future or over some finite stretch of time to reach some target set of outputs. In either case, however, accumulation, the central focus around which other economic questions are grouped, will result from the reinvestment of part of the surplus, and will be analysed either as a case of steady growth or as a deviation from steady growth.

The other approach sees accumulation or decumulation of capital simply as the adjustment of a particular factor of production to its equilibrium level, as determined by supply and demand. In this conception, factor equilibrium is defined in terms of the optimal allocation of scarce resources to competing tasks (in turn defined by the equilibrium final bill of goods, again determined by supply and demand.) The supply of capital may either be taken as given, along with that of land and labour, or it may be seen as governed by saving behaviour, and so responsive to the rate of interest. Demand for capital will be governed by its productivity at the margin, as with the other factors. Equilibrium in a particular sector comes when supply to that sector equals the demand for capital arising in it; equilibrium in general comes when the overall supply of capital equals the overall demand for it. So, according to this conception, accumulation occurs only when the economy is in disequilibrium – it is the movement along the path to equilibrium. The central economic problem is the optimal allocation of scarce resources, and accumulation of capital is a relatively minor matter.

Technical knowledge, however, is itself a scarce resource, and the incentives to produce it and allocate it optimally can be studied by neoclassical methods. Thus the allocation approach can give rise to an account of the long-term transformation of the economy.

But a reallocation process has a natural ending at the equilibrium point, whereas capital accumulation appears to be limitless. Locked into an allocation/disequilibrium framework, the supply and demand approach would be unable to tackle the main questions. It was saved from this fate by the development of the neoclassical growth

model, based on the aggregate production function, and thus combining aspects of the traditional 'surplus' approach with supply and demand. This model provides an account of 'steady growth' over the long run; that is, uniform expansion of all outputs and all inputs, taking place together with regular technical progress. The working of this model, in turn, is based on the traditional theory of competitive factor markets with substitution between labour and capital in the process of production, where both factors are expressed in aggregate terms.

The Keynesian Problematic

The question of substitution initially arose because a simple Keynesian growth model with a given capital–output ratio led to the disturbing conclusion that neither steady growth nor optimal allocation could be achieved. Aggregate demand equals Investment times the multiplier, or I/s, in the simplest case, where s is the average and marginal propensity to save. Aggregate supply, then, is the capital stock times its productivity, or K/v, where v is the capital–output ratio. So the growth rate, $G = I/K = s/v$. This is the rate which equates supply and demand; hence it is the one that business will find satisfactory. But nothing has been said about the labour force or employment; so the equilibrium growth rate need not be consistent with the growth of the labour force, a condition which cannot be optimal. Nor is that the only problem. When I is too low, so that $I/K <$ full employment s/v, $I/s < K/v$, and there will be excess capacity; so businesses will be inclined to reduce I still further. Similarly, when I is too large there will appear to be capacity shortage, and businesses will be inclined to increase I still more. The system gives the wrong signals, and a deviation from steady growth will tend to worsen rather than correct itself.

The Neoclassical Response

Substitution in response to price signals appears to correct this. The neoclassical model determines a path of steady and stable full-employment growth. For instance, when the rate of growth of labour, in efficiency units (the 'natural' rate of growth), persistently exceeds the rate, s/v, determined by the propensity to save and the capital–output ratio (the rate that will just balance aggregate demand and aggregate supply), the real wage will tend to fall, leading firms to substitute labour for capital. As a result, v, the capital–output ratio will decline, raising the rate of growth, s/v. So long as the production function is 'well behaved' (linear and homogeneous, positive first and negative second derivatives, marginal product of capital tends to infinity as K/L tends to zero, and tends to zero as K/L to infinity), there will exist a value of v that will equate s/v to any natural rate of growth. Technical progress which leaves the K/Y ratio unchanged (Harrod-neutral) will not affect the steady-growth path; technical progress which leaves the ratio of the marginal products of capital and labour unchanged (Hicks-neutral) will change the path, but the economy should adjust smoothly to the new equilibrium. In the Keynesian case, investment determined savings; here that causality is reversed (and so the instability disappears – by fiat): in the long run, all savings will be invested; persistent excess capacity (resulting from planned saving $>$ planned investment at full employment) would drive down the rate of interest by lowering the return (or raising the risk) on existing securities; the lower rate of interest will then raise investment up to the full-employment level.

Optimality and the Golden Rule

In neoclassical theory, equilibria tend also to be optimal, but in general the steady growth path will not be. An optimal path ought to be one along which per capita consumption is at a maximum. Consumption is output minus investment, and investment must grow at a fixed rate in order to fully employ the growing labour force. Now consider different capital–output ratios: if the marginal product of capital at a certain v adds more to output than is required to equip the labour force, consumption rises; if it adds less,

consumption falls. Hence when the marginal product of capital just equals the additional investment required for the growing labour force, consumption will be at a maximum. But there is no reason to expect this level of the marginal product to be associated with the capital–output ratio that makes s/v just equal to the rate of growth of the labour force.

The proposition that consumption per head is maximized when the rate of profit equals the rate of growth is sometimes called the 'Golden Rule of Growth'. Under constant returns, it has another disconcerting implication for neoclassical theory. In the stationary state, a positive rate of profit implies that the choice of technique (of the capital–output ratio) is suboptimal. In the stationary state (the normal assumption underlying textbook price theory) only a zero rate of profit is consistent with optimal technique. But a zero rate of profit implies that the Labour Theory of Value governs long-run prices! Either long-run prices are determined by growth theory, or they reflect labour values, or the techniques in use are sub-optimal. (Non-constant returns make this more complicated, but the heart of the problem remains: allocation theory cannot determine long-run prices and optimal techniques independently of growth theory, and therefore of the 'surplus approach'.)

Technical Progress

Treating technical progress as a shift of one kind or another in the production function limits the field of study to changes in method, overlooking the introduction of new products and, indeed, whole new sectors. Treating it as autonomous or as a function of time, even, as in 'learning-by-doing', time on the job, ignores the important influence of demand pressures. Neo-Keynesians, by contrast, treat technical progress as primarily occurring in manufacturing as a response to the growth of demand, so that the rate of technical progress depends on the relative size of manufacturing and on the rate of growth of demand, a relationship known as 'Verdoorn's Law', which has been widely confirmed.

Capital Theory

The standard version of neoclassical theory treats capital as a factor of production, on a par with labour and land, where factors are understood in broad terms and are supplied by households and demanded by firms. (The activity analysis version treats each capital good and each form of land or labour separately, determining its marginal product as a shadow price, thereby avoiding difficulties over capital-in-general, but for that very reason cannot easily analyse the forces that bear on capital as a whole; for instance, saving and investment and their relation to the rate of interest.) The 'surplus' approach of the classics, especially as developed by Marx, conceives capital as an institution: it is a way of organizing production by means of control over produced means of production, which permits processes of production to be valued so they can be bought and sold. These two approaches are obviously different, but are they necessarily incompatible? The capital theory controversy developed over the neoclassical attempt to show that the aggregate production function's implied ordering of techniques (according to an inverse relationship between profitability and capital-intensity) could be constructed in a disaggregated classical or 'surplus' model.

Each point on a neoclassical production function (whether aggregate or not) represents the adoption of a method of production: the firm or the economy as a whole has fully adjusted its plant and equipment. Moving from one point on a production function to another thus means scrapping old plant and replacing it with new, which implies a burst of exceptionally high activity in the capital-goods sector. This will normally be compatible with continuous full employment in the neoclassical framework only if the consumption goods sector is the more capital-intensive, a condition for which there is no economic rationale (Uzawa 1961), or if certain other special conditions are met (Solow 1962). But once we step outside the neoclassical framework the problem of 'traverse' (moving from one growth path to another), even with a *given* technique, can be shown to simply capacity surplus or shortages in one or more sectors, normally accompanied by

temporary overall unemployment (Hicks 1965; Lowe 1976).

In marginal productivity theory a technique is thus uniquely designated by ($K/Y, K/L$); moreover, each K/Y is uniquely paired to its corresponding K/L, and as a direct consequence, each K/L is uniquely associated with a marginal product of capital. But suppose a technique were most profitable at one rate of profit (marginal product of capital) and then also proved the most profitable at another level of the profit rate. If this could happen, the neoclassical production function would not uniquely determine the choice of method of production. Yet the general possibility of this phenomenon ('reswitching') is easily demonstrated. (Not only the neoclassical approach is at risk here; the Marxian doctrine of the falling rate of profit is likewise rendered suspect: Okishio 1962).

Neoclassical production theory, whether aggregate or not, postulates diminishing marginal output as the amount used of a factor is varied in relation to other factors. If factors are paid the value of their marginal products, as the theory of competitive behaviour asserts, then factor reward (e.g. the rate of profit) should fall as the amount of the factor (capital) increases in relation to labour. (If reswitching occurs, it can be demonstrated that at least one of the switches will show a positive relation between capital per worker and the rate of profit.) Once we step outside the conventional approach, this inverse relationship is not intuitively plausible: increasing the amount of capital employed in a production process is a more complex matter than employing more labour. Capital consists of all the various means of production; it is a *set* of inputs. In fact, it is more (and more complicated) than that: at the beginning of production the capital of an enterprise consists of its plant and equipment, its inventory of materials and its wage fund (minus various obligations). A little later it consists of somewhat depreciated plant and equipment, together with the worked-up inventory of marketable goods, while the materials and wage fund have disappeared. But (allowing for changes in indebtedness during production, etc.) although the actual goods in which its capital is embodied are different in the two situations, the business will sell for the same price – it has the same capital value. To vary the amount of capital is to change the size or the nature of the entire process, and it is not at all obvious what effect this will have on the rate of profit.

A second problem concerns influences running the other direction, from the rate of profit to the amount of capital. When the rate of profit changes, competition requires prices to change. (Suppose, *ceteris paribus*, that the real wage rose, requiring the general rate of profit to fall; to keep the rate uniform, so capital will not tend to migrate to the relatively high-profit industries, the prices of labour-intensive products will have to rise relative to capital-intensive ones.) But if the prices of produced means of production change, then the 'amount of capital' embodied in *unchanged* plant and equipment can vary, and this can come about because of variation in the rate of profit. Moreover, the amount of capital embodied in unchanged equipment can vary in either direction when the rate of profit changes, since the direction of relative price changes depends only on relative capital-intensity, about which no general rules can be given. The neoclassical ranking of techniques according to capital-intensity and the rate of return has to be considered an inadequate representation of the real complexities involved in choosing techniques and using capital in production. So the neoclassical answer to the Keynesian problem is not sufficient.

Neo-Keynesian Theory

An alternative to the neoclassical theory of steady growth, however, provides a similar answer by way of a different conception of price adjustments, while still remaining within the conception of accumulation as the expansion, rather than the transformation, of a given system. The overall saving ratio is considered the weighted average of saving out of wages and profits, the weights being the respective income shares. Here the propensity to save out of profits is assumed to be relatively high, and that out of wages to be low. Then, if the natural rate of growth $>$ s/v, eventually the money wage rate would tend to fall, and this, ceteris paribus, would raise the profitability of investment. As a result the overall saving ratio

would rise, bringing s/v up to the full-employment level. If s/v is greater than the natural rate, on the other hand, the resulting excess capacity would lower profitability and tend to bring s/v down. Thus it is not necessary to assume easy and unrealistic substitution; the capital/output ratio can remain fixed, and yet market adjustments will direct the system towards the full employment growth path.

Like the neoclassical, this scenario sees the natural rate of growth as the centre of gravitation towards which the system adjusts. But it has sometimes been given another, more Keynesian interpretation. If, at the level of normal capacity utilization, investment demand were to exceed savings, multiplier pressure would drive up prices – since output could not be (easily) increased. Money wages, on the other hand, would not be driven up, since employment could not be (easily) increased either, for when plant and equipment is operating at full capacity there are no more places on the assembly lines – the full complement of workers has already been hired. Thus the excess demand for goods will *not* translate into excess demand for labour, and prices will be driven up relative to money wages: a Profit Inflation. Thus the overall saving ratio will rise, until the pressure of excess demand is eased. So in the long run as well as in the short, savings adjusts to investment. Understood in this way, the second scenario contradicts the neoclassical one rather than complementing it.

Investment and the Accelerator

But this is still not fully Keynesian, or at least not Harrodian, for the emergence of excess or shortage of capacity must be allowed to influence investment plans – the 'accelerator', or capital-stock adjustment principle. When $s/v >$ actual or current I/K, there will be a slump; when $s/v <$ I/K prices will be bid up relative to money wages. Money wages, in turn, will tend to rise or fall according to whether the actual rate of growth lies above or below the natural. If the actual rate lies above the natural, this will tend to raise the natural and lower the actual. There are thus three rates of growth: the actual, I/K, the warranted, s/v, and the natural, and six possible permutations of these. It can be shown that in only two cases is there an unambiguous tendency for all three rates to converge; in two others, plausible additional assumptions will bring a tendency to converge. But in two cases there seems to be no convergence at all; quite the opposite (Nell 1982). So the Keynesian approach suggests that the full-employment (or, indeed, any) steady growth path should not be treated as a centre of gravitation; it may or may not be what the market tends to bring about.

Capital Value and Profit

Ironically, this neo-Keynesian approach runs afoul of the same problems that plague the neoclassical standard version. For once we leave the one-sector framework, the neo-Keynesian theory implies that excess aggregate demand will bid up, not the price level in general, but the relative price of capital goods – for the excess demand is entirely concentrated in the investment goods sector, and there is no discussion of how this could be transmitted to the consumer-goods sector. Moreover, if both prices did rise relative to money wages, consumer-goods demand would fall. But this would not indicate a possible equilibrium, for it leaves the profit rate unequal in the two sectors. Thus the neo-Keynesian claim must be that a bidding up of the relative price of capital goods will raise the rate of profit, leading to higher savings, etc., but in a two-sector model it is easily seen that this will only be the case when the capital-goods sector is the more capital-intensive. So the validity of the approach depends on an arbitrary condition (which becomes even more arbitrary as the number of sectors increases.)

Even worse, suppose that the capital-goods sector is the more capital-intensive, and consider a small rise in the growth rate to a new equilibrium level, requiring an increased production of capital goods (alternatively, a fall in the actual rate below the equilibrium). The corresponding new overall capital–labour ratio will be higher than the initial one; but to maintain full employment there will have to be a diversion of resources to the industry

with the lower capital–labour ratio. To preserve full employment the capital-goods sector would have to be contracted; but to increase the growth rate it has to expand. (A similar argument holds for a decline in the equilibrium growth rate.) In the case where a rise in the price of capital goods would increase the rate of profit, permitting the neo-Keynesian mechanism to work, the system could not adjust to the new steady growth path, since the two conditions for adjustment contradict one another.

In fact, adjustment from one steady growth path to another turns out to be difficult in general, even without changes in technique. A change in the growth rate requires changes in the relative sizes of sectors, which means shifting labour and resources; but these are normally used in different proportions or in different combinations. And some can only be used in certain sectors and not in others. The 'traverse' from one steady path to another will normally involve both unemployment and shortages, and it may be difficult to actually reach a new path before the conditions determining it change. The 'steady growth' approach to accumulation may face insurmountable problems.

The Significance of Steady Growth

But, what then is the importance of the steady growth path? For the neoclassical approach it is an extension of the concept of equilibrium to the case of expansion over time; for some neo-Keynesians it represented a centre of gravitation, a point towards which the system would move, or around which it would oscillate. For others it may simply be a point of reference – how the system would work if certain contrary-to-fact assumptions held. Real processes will normally be different and can be classified by their distance from such a point of reference.

Following Joan Robinson, steady growth with continuous full employment has been termed a 'golden age'; desired capital accumulation equals the natural rate of growth. But a low desired rate, well below the initial natural rate, might create a large reserve army of unemployed, forcing down real wages and lowering the birth rate, so that the natural rate would fall to the depressed desired rate – a 'leaden age'. A desired rate above the natural rate may bid up real wages enough to lower the rate of profit until the desired rate falls to the natural – a 'restrained golden age'. A 'bastard golden age' occurs when the desired rate cannot be achieved because the real wage cannot be driven down sufficiently, the attempt resulting in inflation. Other possibilities can be envisioned, depending on the adjustment mechanisms postulated. For example, when the initial stock of capital is not appropriate to the desired rate of accumulation, it will have first to be adjusted, but the part of the capital-goods sector that produces capital goods for its own use may be too large or too small for easy adjustment to the desired rate, giving rise to 'platinum age' patterns of accumulation. The catalogue is endless, but its value is limited.

Steady growth, in fact, appears to be best analysed as a supply-side concept. Its most elaborate development, in fact, is strictly supply-side – as the von Neumann ray, or in Sraffa's terms, the Standard Commodity, where the industry sizes of the system have been so adjusted that the net product of the economy as a whole consists of the same commodities in the same ratios as its aggregate means of production. The warranted rate of growth, by contrast, balances supply and demand. But it is an imperfect growth concept, for it balances aggregate supply and aggregate demand *at a moment of time*; it does not balance the growth of supply with the growth of demand. The von Neumann ray is an analysis of the growth of supply – but so far there is no comparably detailed analysis of the growth of demand.

Accumulation and Technical Change

This not only brings to light a defect in the theory of steady growth, it also raises the question of the relation of steady growth to the accumulation of capital. For the best-established empirical proposition in the study of consumer behaviour states that as income increases, consumer demand will increase non-proportionally – it will shift in a characteristic manner. Hence there is little point in trying to complete the theory of steady growth

with an account of steady growth in demand; it doesn't happen.

In actual fact, steady growth has never taken place. The history of capitalism is a history of successive booms and slumps, but perhaps even more striking, of slow but persistent long-term shifts in crucial relationships. For two centuries labour shifted out of agriculture and migrated to the cities to work in manufacturing industry. For over half a century now labour has shifted into services, first from agriculture and then, later, from manufacturing as well. For almost a century the relative size of the government sector has been rising, whether measured by share of GNP or by share of employment.

These points lead to a major criticism of the treatment of technical progress in accumulation: whether it is presented as shifting the production function, as learning by doing, or in a 'technical progress' function, and whether conceived as embodied or disembodied, it has been treated as leading to the extraction of greater output from given resources, in the context of steady growth. But technical progress introduces new products as well as new processes, and together these change the forms of social life. This is reflected in the changing importance of the major sectors of the economy, in the changing class structure and in the changing patterns and nature of work. None of these points seems to be captured by the current analyses, in part because of the preoccupation with steady growth, based on an overly simplified concept of capital as productive goods. When capital is understood as also being a form of organization, then the link between accumulation and the transformation of institutions can be forged. Another reason, perhaps, may be that technical progress has been approached too timidly, and without understanding its dual relation to the growth of demand. For technical progress both stimulates the growth of demand and responds to it.

Steady Growth Versus Transformational Growth

In practice, steady growth is an impossibility for at least three reasons. First, land and natural resources are limited, and high-grade ores and high-fertility lands are the first to be used. As they are used up over time, productivity falls unless and until technical progress offsets the decline – but such technical progress will have to involve new products. Second, as mentioned, Engel curves imply that consumption patterns will be changing. And finally, if propensities to save differ in the different social classes (and if workers receive interest on their savings, and capitalists salaries for managing capital), then the relative wealth of the classes will be changing over time, leading to changes in the composition of demand. The first point implies that costs will tend to rise; the second two, that demand for consumer goods will tend to rise more slowly as time passes. All three therefore point to long-term stagnation in the absence of major technological changes.

This does not simply mean increasing the productivity of currently employed processes; it means the development of new processes and new products – both for consumers and for industry. It means electrification, or the internal combustion engine, the aeroplane or, perhaps, the computer. The changes must be of sufficient importance to lead to an investment boom resulting from widespread scrapping of present plant and equipment, as well as the development, concurrently, of large-scale new markets, as consumers introduce the new products into their living patterns. And as new plants are built, economies of scale can be realized, making it possible to lower prices, so as to reach new markets in lower levels of the income distribution. Capital organizes markets and marketing as well as production.

New household products have emerged because a way has been found to perform some normal daily activity better or more cheaply by, in effect, shifting it from the household to industry, capitalizing it. New industrial processes, usually involving new products as well, have emerged as the result of mechanizing activities formerly performed by workers, enabling them to be done better, or more cheaply, or more reliably. Mass-production goods have replaced home crafts; the mechanization of agriculture, in conjunction with

Engel's Law, has displaced farm labour; the rise of manufacturing, to build the factories and then to supply the new goods, has provided employment for the displaced labour – but at greatly reduced hours of work per week, providing more hours to spend on consuming.

The rise of mass production and the consequent urbanization have created new problems; among others, periodic mass unemployment, which in turn had to be dealt with by an expanded government. And today traditional mass production is being transformed by the computer and the chip, with consequences we cannot yet fully foresee.

The interlocking emergence of new products and new processes, creating new markets and new industries, can be termed 'transformational growth', in contrast to steady growth. It is here that the true story of the accumulation of capital, and the causes of the wealth of nations, will be found, but to date this study has been left the province of economic historians.

See Also

- ▶ Classical Growth Models
- ▶ Neoclassical Growth Theory

References

Domar, E.D. 1946. Capital expansion, rate of growth and employment. *Econometrica* 14: 137–147.

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49: 14–33.

Hicks, J. 1965. *Capital and growth*. Oxford: Oxford University Press.

Kaldor, N. 1956. Alternative theories of distribution. *Review of Economic Studies* 23: 83–100.

Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.

Laibman, D., and E.J. Nell. 1977. Reswitching, Wickell effects and the neo-classical production function. *American Economic Review* 67: 878–888.

Lowe, A. 1976. *The path of economic growth*. New York: Cambridge University Press.

Marx, K. 1867–1894. *Capital*, vols. I, II, III. Moscow: Progress Publishers, n.d.

Nell, E.J. 1982. Growth, distribution and inflation. *Journal of Post-Keynesian Economics* 5: 104–113.

Nell, E.J. 1986. *Priority and public spending*. London: George Allen & Unwin.

Ricardo, D. 1817. *On the principles of political economy and taxation*. Vol. I of *The works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Robinson, J. 1956. *The accumulation of capital*. London: Macmillan.

Robinson, J. 1962. *Essays in the theory of economic growth*. London: Macmillan.

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94.

Solow, R. 1962. Substitution and fixed proportions in the theory of capital. *Review of Economic Studies* 29: 207–218.

Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Uzawa, H. 1961. On a two-sector model of economic growth. *Review of Economic Studies* 24: 40–47.

Von Neumann, J. 1937. A model of general economic equilibrium. *Review of Economic Studies* 13(1945–6): 1–9.

Acyclicity

Douglas Blair

Acyclicity is a consistency property of preferences and other binary relations. It requires that the asymmetric part P of the relation (e.g. the subrelation of strict preference) contain no cycles; that is, for no sequence of alternatives $x_1, x_2, \ldots, x_n$ is it true that $x_1 P x_2$, $x_2 P x_3, \ldots, x_{n-1} P x_n$, and $x_n P x_1$. The study of cyclic preferences dates at least to Condorcet's (1785) treatment of the paradox of voting, in which transitive individual voters generate cyclic majority preferences.

Whenever a feasible set S contains more than two alternatives, some principle is needed to generate choices $C(S)$ from the pairwise comparisons summarized by the preference relation; one natural candidate is the set of undominated alternatives. Acyclicity is necessary and sufficient for the existence of a non-empty set of undominated elements in any finite feasible subset S of the universal set of alternatives. In addition, defining the choice set as the undominated alternatives according to an acyclic relation guarantees that choices will exhibit a desirable consistency

property: if S is a subset of T and if x belongs both to S and to $C(T)$, then x must belong to $C(S)$. In Sen's (1970) example, if the world champion is a Pakistani, then he must be champion of Pakistan as well. This property is attractive in piecemeal decision mechanisms in which choices are made from unions of choices over subsets. If an alternative fails to be chosen in some subset, it need not be reconsidered later, since the contrapositive of this property ensures that the alternative will not be among the final choices.

Acyclicity is a significantly weaker consistency property than transitivity; it permits intransitivities both of the strict preference relation P and the symmetric subrelation of indifference I. For example, the preferences xPy, yPz, and xIz are acyclic; so too are the preferences xIy, xIz, and xPz.

Acyclicity arises in several contexts in economics. Consumer theory's Strong Axiom of Revealed Preference (see Houthakker 1950; Ville 1951–2), for example, is an axiom asserting that a particular revealed preference relation is acyclic. It arose as well early in the development of game theory; the acyclicity of dominance relations is closely linked to the uniqueness of the von Neumann–Morgenstern (1947, ch. XII) solution.

Acyclicity has been studied most intensively, however, in connection with Arrow's (1951) Impossibility Theorem. This proposition concerns constitutions, which aggregate sets of individuals' preference orderings into social preferences. Arrow showed that the only constitutions satisfying two reasonable axioms and yielding transitive social preferences are dictatorial. Several writers have attempted to circumvent this conclusion by relaxing transitivity to the more defensible requirement of acyclicity. Non-dictatorial acyclic constitutions do exist, but they turn out to be hardly more attractive than dictatorships. Blair and Pollak (1982) review this literature and show that such constitutions must endow at least one voter with extensive veto power over strict social preferences opposite his or her own. If egalitarian concerns force the vesting of such power in many such voters, the constitution will be highly indecisive, that is, frequently yield judgements of indifference between alternatives.

See Also

- ▶ Arrow's Theorem
- ▶ Orderings
- ▶ Preferences
- ▶ Preordering
- ▶ Revealed Preference Theory
- ▶ Social Choice
- ▶ Transitivity

References

Arrow, K. 1951. *Social choice and individual values*. New York: Wiley.
Blair, D.H., and R. Pollak. 1982. Acyclic collective choice rules. *Econometrica* 50(4): 931–943.
Condorcet, Marquis de. 1785. *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix*. Paris.
Houthakker, H. 1950. Revealed preference and the utility function. *Economica* 17: 159–174.
Sen, A. 1970. *Collective choice and social welfare*. San Francisco: Holden-Day.
Ville, J. 1952. The existence conditions of a total utility function. *Review of Economic Studies* 19(2): 123–128.
von Neumann, J., and O. Morgenstern. 1947. *Theory of games and economic behavior*, 2nd ed. Princeton: Princeton University Press.

Adams, Henry Carter (1851–1921)

A. W. Coats

Adams was born on 31 December 1851 in Davenport, Iowa, and died in Ann Arbor, Michigan, on 11 August 1921. In many respects typical of the new generation of late nineteenth-century American social scientists, Adams became a professional economist only after considering a career in the church or in reform political journalism. After graduating from Iowa (later Grinnell) College in 1874, he spent 1 year as a school teacher and another studying at Andover Theological Seminary before obtaining a fellowship at the newly founded Johns Hopkins University, where he received its first PhD, in 1878.

At Hopkins, Francis Walker steered him towards public finance, a field to which Adams subsequently made major pioneering contributions. But he was no narrow specialist, and 2 years' further study in Europe, mainly at Berlin and Heidelberg, laid the foundations for the breadth of interest, historical perspective, and philosophical insight that characterized his later writings.

On returning to the USA Adams, like many of his contemporaries, found difficulty in obtaining a satisfactory permanent academic post and was obliged to spend several years in temporary or part-time employment before obtaining a permanent position at the University of Michigan in 1886, where he spent the remainder of his career. The frank and revealing correspondence between Adams and President Angell immediately prior to this appointment is a notable contribution to the chequered history of academic freedom in America (cf. Dorfman 1954, editor's introduction; Coats 1968), for Adams had only recently been dismissed from Cornell for having publicly expressed support for labour unions during the outcry over the Haymarket bomb incident. At Ann Arbor, Adams built up a distinguished department (Brazer 1982) and achieved national recognition for his nearly two decades of service as Chief Statistician to the Interstate Commerce Commission, where by constructing and implementing a system of uniform railway accounts he made a lasting contribution to the development of public regulation.

A co-founder and staunch supporter of the American Economic Association, of which he was President in 1896–1897, Adams endeavoured to bring the best elements in European economic, social, and political thought to bear on the study of contemporary problems. He made no significant contributions to economic theory, although he was one of the first American economists to incorporate Jevons's value theory into his teaching. A more temperate critic of laissez-faire individualism than Richard T. Ely, Adams preferred clear thinking to exhortation and was respected by his peers for his solid scholarship and balanced judgement, for example in his seminal essay on the 'Relation of the State to Industrial Action', the first systematic American examination of the respective spheres of private and public economic activity. While recognizing the force of competition as a principle he considered it inadequate as a curb to monopoly power, and liable to depress the ethical plane of economic activity as unscrupulous employers undercut their more reputable rivals. Arguing the need for increased government intervention as the economic system became more complex, Adams nevertheless opposed socialism and nationalization, initially preferring municipal and state to federal regulation. Later he viewed the regulatory commission as the ideal conservative instrument of reform. His analysis of the distinction between increasing, constant and diminishing returns underlay his concern at the growth of corporate power and at the end of his life he advocated cooperation as the most desirable basis for industrial reform. Like many later thinkers he emphasized the need for collaboration between the various organized groups in society, and his emphasis on the worker's proprietary rights in his employment became a significant theme in the writings of American labour economists. Another pioneering contribution was his appreciation of the interdependence of economics and jurisprudence, one of many elements drawn from the tradition of German historical economics. Although he displayed little interest in the monetary questions which troubled so many of his contemporaries, Adams was a versatile and fertile thinker, many of whose ideas became common currency among later generations of American social scientists.

Selected Works

1881. *Outline of lectures upon political economy*. Amherst: C.A. Bangs. 2nd ed., Ann Arbor, 1886.
1884. *Taxation in the United States 1789–1816*. Baltimore: Johns Hopkins Press.
1887a. *Public debts: An essay in the science of finance*. New York: D. Appleton. 2nd ed., 1898.
1887b. *Relation of the state to industrial action*. Baltimore: American Economic Association. New ed., ed. J. Dorfman, New York: Columbia University Press, 1954.

1897. *Economics and jurisprudence*. London: Macmillan; New York: S. Sonnenschein. New ed., ed. J. Dorfman, New York: Columbia University Press, 1954.
1898. *The science of finance: An investigation of public expenditures and public revenues*. New York: H. Holt. Rev. ed., 1924.
1918. *American Railway Accounting: A commentary*. New York: H. Holt.

References

Brazer, M.C. 1982. The Economics Department at the University of Michigan: A centennial perspective. In *Economics and the world around it*, ed. S.H. Hymans. Ann Arbor: University of Michigan Press.
Coats, A.W. 1968. Henry Carter Adams: A case study in the emergence of the social sciences in the United States, 1850–1900. *Journal of American Studies* 2: 177–197.

Adaptive Estimation

Douglas G. Steigerwald

Abstract

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

Keywords

Adaptive estimation; Kernel estimator; Linearized likelihood estimation; Maximum likelihood; Nonparametric estimation; Semiparametric estimation; Spline functions

JEL Classifications

C14

An adaptive estimator is an efficient estimator for a model that is only partially specified.

For example, consider estimating a parameter that describes a sample of observations drawn from a distribution F. One natural question is: is it possible that an estimator of the parameter constructed without knowledge of F could be as efficient (asymptotically) as any well-behaved estimator that relies on knowledge of F? For some problems the answer is 'yes', and the estimator that is efficient is termed an adaptive estimator.

Consider the familiar scalar linear regression model (in which we let t rather than i index observations)

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

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

The question then is: how can we verify that an estimator is adaptive? As there will generally be

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

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

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

While Stein's condition has intuitive appeal, it is not straightforward how to use the condition to define estimators that are adaptive. In an invited lecture, Bickel (1982) laid out a simpler condition that does yield a straightforward link to the construction of adaptive estimators. To understand the condition, let E_F denote expectation with respect to the population error distribution and let $E_{\tilde{F}}$ denote expectation with respect to an arbitrary distribution $\tilde{F} \in \mathscr{F}$. Let l be the log-likelihood for the regression model with data $z = (y, x)$ and let $\dot{l}(z, \beta, F)$ denote the score for β, constructed from the model in which F is the error distribution. A familiar condition that arises in the context of likelihood estimation is that the expected population score $E_F[\dot{l}(z, \beta, F)]$ equal 0. Bickel's condition is simply that the population score must have expectation zero over the entire family $\mathscr{F}$, that is, for any $\tilde{F} \in \mathscr{F}$,

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

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

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

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

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

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

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

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

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

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

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

$$dP_t = m_t dt + \upsilon_t dB_t.$$

The presence of standard Brownian motion, B_t, makes the model of price a stochastic differential equation. The function m_t captures the deterministic movement or drift while υ_t is the potentially time-varying scale of the random component. Lepski and Spokoiny (1997) study the model in which υ_t is constant and m_t is unknown. They establish that a nonparametric estimator of m is pointwise adaptive. Yet an estimator that is pointwise adaptive – that is, for a given point t_0 the nonparametric estimator of $m(t_0)$ is asymptotically efficient – may not perform well for all values within the range of the function m. Such an idea is intuitive; without knowledge of the smoothness of m, estimators designed to be optimal for one value of t may be very different from optimal estimators for another value of t. Cai and Low (2005) study efficient estimation of m over neighbourhoods of t_0 and show that an estimator constructed from wavelets is adaptive. The restriction that the scale is constant is often difficult to support with financial data. A more realistic model, which Mercurio and Spokoiny (2004) study, models the asset return as a stochastic differential equation with drift 0 and v_t varying over time. The time-varying scale is assumed to be constant over (short) intervals of time, but is otherwise unspecified. They construct a nonparametric estimator of the volatility from a kernel that performs local averaging and show that the resultant estimator is adaptive.

See Also

- Efficiency Bounds
- Partial Linear Model
- Semiparametric Estimation

Bibliography

Bickel, P. 1982. On adaptive estimation. *Annals of Statistics* 10: 647–671.

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

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

Faraway, J. 1992. Smoothing in adaptive estimation. *Annals of Statistics* 20: 414–427.

Lepski, O., and V. Spokoiny. 1997. Optimal pointwise adaptive methods in nonparametric estimation. *Annals of Statistics* 25: 2512–2546.

Mercurio, D., and V. Spokoiny. 2004. Statistical inference for time-inhomogeneous volatility models. *Annals of Statistics* 32: 577–602.
Newey, W. 1988. Adaptive estimation of regression models via moment restrictions. *Journal of Econometrics* 38: 301–339.
Portnoy, S., and R. Koenker. 1989. Adaptive L-estimation for linear models. *Annals of Statistics* 17: 362–381.
Steigerwald, D. 1992. On the finite sample behavior of adaptive estimators. *Journal of Econometrics* 54: 371–400.
Stein, C. 1956. Efficient nonparametric testing and estimation. In *Proceedings of the third Berkeley Symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley: University of California Press.

Adaptive Expectations

Michael Parkin

JEL Classifications
E17

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

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

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

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

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

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

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

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

Second, the adaptive expectations hypothesis was empirically easy to employ. Koyck (1954) showed how a simple transformation of an equation with an unobservable expectation variable in

it could be rendered observable by performing what became a famous transformation bearing Koyck's name. If some variable, y, is determined by the expected future value of x, that is:

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

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

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

An equation like this may be used to estimate not only the desired values of α and β but also the value of λ, the coefficient of expectations adjustment. Thus, economists seemed to have a very powerful way of modelling situations in which unobservable expectational variables were important and of discovering speeds of response both of expectations to past events and of current events to expectations of future events.

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

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

See Also

- ▶ Cobweb Theorem
- ▶ Expectations
- ▶ Hyperinflation
- ▶ Phillips Curve
- ▶ Rational Expectations

Bibliography

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

Fisher, I. 1911. *The purchasing power of money*. New York: Macmillan. (latest edition, A.M. Kelley, New York, 1963).

Friedman, M. 1957. *A theory of the consumption function*. Princeton: Princeton University Press.

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

Koyck, L.M. 1954. *Distributed lags and investment analysis*. Amsterdam: North-Holland.

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

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

Mincer, J. 1969. *Economic forecasts and expectations*. New York: National Bureau of Economic Research.
Muth, J.F. 1960. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association* 55: 299–306.
Nerlove, M. 1958. *The dynamics of supply: Estimation of farmers' response to price*. Baltimore: Johns Hopkins Press.
Phelps, E.S. 1970. *Microeconomic foundations of employment and inflation theory*. New York: Norton.

Addiction

George Loewenstein and Scott Rick

Abstract

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

Keywords

Addiction, rational account of; Becker, G.; Drugs; Excise tax; Forward price elasticity; Rational behaviour; Sin taxes; Time consistency

JEL Classifications

D1

Economists were latecomers to the study of addiction, a concept which researchers in other disciplines usually define as including a loss of self-control, continuation of behaviour despite adverse consequences, and preoccupation or obsession with the substance or activity one is addicted to. Economists came late to the subject perhaps because the first two of these characteristics seem inconsistent with economists' rational choice paradigm.

This may be exactly what spurred Gary Becker, along with coauthor Kevin Murphy, to propose, in 1988, a 'rational account of addiction', which stimulated much subsequent research and theorizing by economists. Although not the first economic account of addiction, Becker and Murphy's model (referred to henceforth as B&M) was certainly the most influential, and has spawned a very lively line of research, theorizing and debate about addiction by economists.

Contributions of Disciplines Other than Economics

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

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

In addition to generating a wide range of interesting and important findings, researchers in disciplines other than economics have proposed a variety of theoretical perspectives on addiction. Some perspectives place great importance on the pleasure of drug-taking, the pain of withdrawal, or the motivational force of 'cue-conditioned' craving, while others view drug use as a form of self-medication for psychiatric conditions such as depression.

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

Becker and Murphy's Model

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

This model can accommodate a number of features of classical addiction, such as that being addicted lowers instantaneous utility $u_s < 0$, that it increases the instantaneous marginal utility of taking the drug $u_{cS} > 0$. Solving the model yields

a number of implications, most importantly that it can be rational for an individual to maintain a positive rate of consumption of an addictive good.

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

Moving Beyond Becker and Murphy

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

One important generalization has been to examine the implications of relaxing the assumption of exponential time discounting. Gruber and Koszegi (2001, 2004), for example, propose a model in which time-inconsistent addicts have self-control problems: they would like to quit using but cannot force themselves to do so (see also O'Donoghue and Rabin 1997). As in B&M, Gruber and Koszegi's model predicts that a rise in current or anticipated excise taxes will reduce use of addictive substances. However, although the models make similar behavioural predictions, they interpret the hedonic consequences of altered usage behaviour differently. B&M predicts that taxes on addictive substances – 'sin taxes' – make addicts worse off since the price of a good that they enjoy has risen. Gruber and Koszegi's model, on the other hand, predicts that the tax makes time-inconsistent addicts better off since it provides a valuable self-control device.

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

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

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

Finally, a recent line of theoretical models, while also building on the insights of Becker and Murphy, has incorporated evidence from the psychological literature on cue-conditioned craving

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

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

See Also

- ▶ Becker, Gary S. (Born 1930)
- ▶ Intertemporal Choice
- ▶ Rational Behaviour

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

Bibliography

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

Bernheim, B., and A. Rangel. 2004. Addiction and cue-triggered decision processes. *American Economic Review* 94: 1558–1590.

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

Gruber, J., and B. Koszegi. 2001. Is addiction 'rational?' Theory and evidence. *Quarterly Journal of Economics* 116: 1261–1305.

Gruber, J., and B. Koszegi. 2004. A theory of government regulation of addictive bads: Tax levels and tax incidence for cigarette excise taxation. *Journal of Public Economics* 88: 1959–1987.

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

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

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

Lyvers, M. 2000. 'Loss of control' in alcoholism and drug addiction: a neuroscientific interpretation. *Experimental and Clinical Psychopharmacology* 8: 225–245.

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

Pacula, R., and F. Chaloupka. 2001. The effects of macro-level interventions on addictive behavior. *Substance Use and Misuse* 36: 1901–1922.

Robins, L. 1973. *The Vietnam drug user returns*. Washington, DC: U.S. Government Printing Office.

Slovic, P. 2000a. What does it mean to know a cumulative risk? Adolescents' perceptions of short-term and long-term consequences of smoking. *Journal of Behavioral Decision Making* 13: 259–266.

Slovic, P. 2000b. Rejoinder: The perils of Viscusi's analyses of smoking risk perceptions. *Journal of Behavioral Decision Making* 13: 273–276.

Viscusi, W. 2000. Comment: The perils of qualitative smoking risk measures. *Journal of Behavioral Decision Making* 13: 267–271.

Wertenbroch, K. 1998. Consumption self-control via purchase quantity rationing of virtue and vice. *Marketing Science* 17: 317–337.

Wertenbroch, K. 2003. Self-rationing: Self-control in consumer choice. In *Time and decision: Economic and psychological perspectives on intertemporal choice*, ed. G. Loewenstein, D. Read, and R. Baumeister. New York: Russell Sage Foundation.

Adding-Up Problem

Ian Steedman

In any theory of income distribution in which one type of return is determined residually, it will be

tautologically true that the various different incomes, as determined by the theory, will add up so as to exhaust the total product. By contrast, any theory which provides a 'positive' explanation for every category or return, treating none as a residual, must show that the various returns so explained do indeed exhaust the product. In practice, it has been with reference to the marginal productivity theory that this consistency requirement has received considerable attention. By the early 1890s a number of authors had sought to extend the 'principle of rent' into a completely general theory of distribution but it was P.H. Wicksteed, in his Co-ordination of the Laws of Distribution (1894) who first clearly stated, and attempted to resolve, the resulting adding-up problem.

Consider first the simplest case, in which all markets are perfectly competitive, there is no uncertainty and 'entrepreneurs' are seen as mere hiring agents. If it is supposed also that all productive processes exhibit constant returns to scale, then the adding-up problem is shown by Euler's theorem on homogeneous functions to be a quite trivial problem, as Flux (1894) pointed out in his *Economic Journal* review of Wicksteed's book. When assuming constant returns one should, of course, be mindful of Samuelson's warning that 'Any function whatever in n variables may be regarded as a subset of a larger function in more than n variables which is homogeneous of the first order' (1983, p. 84, n. 13). Attention can also be drawn to the indeterminacy of the sizes of firms in the constant returns case, and thus to the question of how the perfect competition assumption can be underpinned, but these (perfectly proper) questions are not specific to the adding-up problem. It is, however, vital to appreciate that linear homogeneity of production relations does *not*, by itself, dispose of the adding-up problem; it is linear homogeneity in production, combined with perfectly competitive market conditions, which does that. This was forcefully demonstrated by Wicksteed himself in 1894. Whilst he upheld the assumption of constant returns to scale in production, he also held that a proportional increase of *all* inputs – both those used in production and those used in selling activities – would not result in an equal proportional increase in the quantity sold, at a given price. Thus there is not a 'constant returns' relationship between total outlays and total revenue. Wicksteed examined the consequences for 'adding-up', first in the case of monopoly and then with an ever-increasing number of firms in the industry, and was able to show that, as the number of firms became very large, marginal productivity pricing would approximately exhaust the product. Adding-up, or otherwise, is thus intimately related to market conditions.

Wicksteed's assumption of linear homogeneity in production, together with what was taken by Walras, at least, to be his implicit slighting of the work of others, resulted in his work receiving a hostile response from Pareto, Edgeworth and Walras. In the third edition of his *Eléments* Walras inserted an Appendix III, dated October 1895, which ended with the words 'Mr. Wicksteed . . . would have been better inspired if he had not made such efforts to appear ignorant of the work of his predecessors'. (This appendix was, however, dropped from subsequent editions; Stigler and Schumpeter have disagreed over the precise import of, and degree of justification for, Walras' displeasure.) More constructively, the second half of Walras' appendix outlined a proof of the adding-up theorem under competitive conditions (see below), a proof based on work by Barone. (It seems that Barone had submitted a review of Wicksteed's book to the *Economic Journal* and that Edgeworth had first accepted the review for publication but then subsequently withdrew his acceptance.) In his *Economic Journal* (1906) review of Pareto's *Manuale di Economia Politica* (1906), Wicksteed acknowledged the justice of the criticisms which Edgeworth and Pareto had made of his 1894 *Co-ordination* argument; and in the *Common Sense* (1910) he again referred to Edgeworth and Pareto and stated that paragraph 6 of the *Co-ordination* 'must be regarded as formally withdrawn' (p. 373, n. 1). (It is to be noted that Wicksteed does *not* refer to Walras in either of these acknowledgements of justified criticism.) In Volume I of his *Lectures on Political Economy* (1901), Wicksell expressed surprise that

Wicksteed had 'declared – for reasons difficult to understand – that he desired to withdraw this work [the *Co-ordination*]' (1934, p. 101, n. 4). It must be noted clearly, first that Wicksteed did *not* withdraw the work as a whole, but only its paragraph 6, and secondly that Wicksteed's proof of the adding-up theorem under linear homogeneity and perfect competition is contained in paragraph 5. Paragraph 6, which he did declare to be withdrawn, concerns the extension of the result of paragraph 5 to the cases of imperfect product markets and of more than two inputs. This, together with Wicksteed's continued use of marginal productivity theory in his *Common Sense*, supports the view of Hutchison, Robbins and Stigler that Wicksteed's 'recantation' was 'merely verbal', and not a rejection of the substance of his earlier argument.

The solution to the adding-up problem which can be associated with the names of Barone, Walras and Wicksell dispenses with the linear homogeneity assumption but is still concerned with long run perfectly competitive equilibrium; it is centred not on the industry but on the individual firm. Any cost minimizing firm, which faces diminishing marginal products and given input prices, will so arrange its production that $w_i = (\text{mc})(\partial q/\partial x_i)$, for *each i*, where w_i is the price of the *i*th variable input, $(\partial q/\partial x_{\text{i}})$ its marginal product, and (mc) the marginal cost of the output in question. Multiplying both sides by x_i and then summing over *i*, one finds that $(\text{avc})q = (\text{mc})\sum x_i(\partial q/\partial x_i)$, where (avc) is average variable cost and q is output. For the cost minimizing firm, then, $\sum x_i(\partial q/\partial x_i) \gtreqless q$ according as $(\text{mc}) \gtreqless (\text{avc})$, that is according as average variable cost is falling, constant, or rising. If the average variable cost curve has a minimum point then, at that point, it will be as *if* there are constant returns to scale and 'adding-up' will obtain. Now introduce the assumption of profit maximization; the perfectly competitive firm will obey the rule $p = (\text{mc}) \geq (\text{avc})$, where p is the product price. Hence $\sum x_i(\partial q/\partial x_i) \leq q$ for such a firm – and equality will hold in, and only in, the long-run equilibrium position (with p = (mc) = (avc) = minimum average total cost).

Consider now the long-run equilibrium position under imperfect competition. The results given above for the cost minimizing firm will still hold, of course, but now (mc) is equal to marginal revenue rather than to product price. The consequence is that, in an 'imperfect' long run equilibrium, $\sum x_i(\partial q/\partial x_i) = (e/e-1)q$, where '$e$' is the (absolute) elasticity of the demand curve at the equilibrium point. (This result naturally tends to the corresponding perfectly competitive result as 'e' tends to infinity.) Analogous but inevitably more complex results can, of course, be obtained when both product and input markets are imperfect.

In the subsequently withdrawn paragraph 6 of his *Co-ordination*, Wicksteed noted that 'In practical cases there is usually a speculator who . . . buys the other factors, speculatively, at their *estimated* values' (p. 41, emphasis added) and that the speculator may make a gain or a loss, depending on how those anticipated values compare with the actual, realized values. He continued: 'But these gains and losses may be resolved into (1st) compensation for risk, and (2nd) the share that falls to this special speculating ability, regarded as a factor of production, and receiving its share of the production in accordance with the general formula [of marginal productivity]' (p. 42). Can entrepreneurship properly be regarded as simply 'another factor'? If not – and Edgeworth and Wicksell, for example, appear to have thought not – if entrepreneurship is related to true uncertainty (as opposed to risk) and if uncertainty leads to the existence of *residual* 'pure profits' then, as observed above, there is no 'adding-up problem' to be solved. For that problem arises, within the marginal productivity context, only when *every* form of income is related to the marginal product of some input.

See Also

- ▶ Euler's Theorem
- ▶ Marginal Productivity Theory
- ▶ Wicksteed, Philip Henry (1844–1927)

References

Edgeworth, F.Y. 1904. The theory of distribution. *Quarterly Journal of Economics*. Reprinted in F.Y. Edgeworth, *Papers relating to political economy*, vol. I. London: Macmillan, 1925.

Flux, A.W. 1894. Review of Wicksell's Über Wert, Kapital und Rente and of Wicksteed's co-ordination, etc. *Economic Journal* 4: 305–313.

Hicks, J.R. 1968. *The theory of wages*, 2nd ed. London: Macmillan.

Hutchison, T.W. 1953. *A review of economic doctrines, 1870–1929*. Oxford: Clarendon Press. ch. 5.

Robbins, L. 1933. Editorial introduction to P.H. Wicksteed. *The common sense of political economy*. London: Routledge & Kegan Paul.

Robinson, J.V. 1934. Euler's theorem and the problem of distribution, *Economic Journal*. Reprinted in *Collected economic papers of Joan Robinson*, vol. I. Oxford: Blackwell, 1966.

Samuelson, P.A. 1983. *Foundations of economic analysis*, enlargedth ed. Cambridge, MA: Harvard University Press. ch. 4.

Stigler, G.J. 1968. *Production and distribution theories: The formative period*. New York: Agathon Press. ch. 12.

Walras, L. 1874–7. *Elements of pure economics*. London: Allen & Unwin, 1954 (Appendix III).

Wicksell, K. 1934. *Lectures on political economy*, vol. I. London: Routledge & Kegan Paul. esp. pp. 101, 125–33.

Wicksteed, P.H. 1894. *An essay on the co-ordination of the laws of distribution*. London: Macmillan. Subsequent editions, ed. L. Robbins, LSE Reprints, 1932; ed. I. Steedman. London: Duckworth, 1987.

Adjustment Costs

Aubhik Khan and Julia K. Thomas

Abstract

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

Keywords

Adjustment costs; Adjustment hazards; Aggregate nonlinearities; Business cycles; Capital irreversibility; Convex cost functions; Distributed lags; Dynamic stochastic equilibrium analysis; Euler equations; Equilibrium; Frictions; Intermediate goods; Inventory policies; Investment theory; Linear quadratic inventory models; Lumpy investment; Market-clearing relative prices; Monetary non-neutralities; Neoclassical investment theory; Nonlinear microeconomic adjustment; Partial adjustment; Piecewise-linear adjustment costs; Production functions; Quadratic cost functions; Rational expectations; (S,s) decision rule; (S, s) policies; Tobin's q; Total factor productivity

JEL Classifications

D4; D10

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

In these few pages, we describe the mechanics of commonly used adjustment costs and briefly discuss their role in several leading macroeconomic applications. Since a comprehensive

survey is beyond the scope of this article, many important applications have been excluded. However, where possible we direct the reader to influential research on these topics.

Convex Costs

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

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

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

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

Without some theoretical basis to explain their empirical success, partial adjustment models might have been abandoned quickly. A partial resolution arrived in the mid- to late 1960s with the application of capital adjustment costs in models of investment (see Eisner and Strotz 1963; Lucas 1967a, b; Gould 1968; Treadway 1971). There, gradual aggregate adjustment broadly consistent with the analogue to (1) was obtained by assuming that, beyond other costs associated with the acquisition of capital (for example, user costs), the very act of adjusting the capital stock incurred real output costs. These costs, $\Phi(k', k)$, were strictly increasing and convex in the distance between the chosen new level of capital and the current level, $|k' - k|$, thereby implying a smoothly rising marginal adjustment cost in the size of the current adjustment. As such, they introduced dynamic elements into the firm's previously static decision problem and led it to smooth its investment activities over time. Nonetheless, so long as the treatment of expectations was incomplete, the mapping to a partial adjustment equation could not be robustly established.

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

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

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

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

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

For researchers implementing equations like (1), an important contribution of Sargent's model was in illustrating how the very features that linked current employment to its lagged determinants also necessarily divorced each date's target, N_t^*, from the statically derived optima assumed in early partial adjustment estimations. Notice that the firm's target in (2) involves expectations of each variable affecting the future value marginal product of labour, because, given adjustment costs, this current choice influences its future level of employment. Moreover, as an increase in the adjustment cost parameter, φ, shifts the marginal adjustment cost schedule upward at all dates, it not only implies a slower adjustment rate (lower λ) but also increases the influence of these expectations of future variables in the determination of the current target.

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

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

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

Non-convex Costs

Despite their relative success in reproducing the persistence of aggregate series, empirical models based on convex adjustment costs have fared poorly along other dimensions. For example, estimations of the neoclassical investment model attribute very low explanatory power to average Q and assign large coefficients to adjustment cost parameters in explaining changes in investment (Chirinko 1993; Caballero 1999). Large estimates of adjustment costs, which in turn imply

implausibly slow adjustment speeds, are also a recurring problem for linear quadratic inventory models (Ramey and West 1999). Elsewhere, the sharp difference between rates of employment adjustment estimated from high-frequency firm-level data and those estimated from low-frequency aggregate data suggests spatial and temporal bias inconsistent with the common assumption of symmetric quadratic adjustment costs (Hamermesh and Pfann 1996). Moreover, there is mounting microeconomic evidence suggesting that the predominant adjustment frictions confronting firms in actual economies may be non-convex, rather than convex, in nature.

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

(S, s) Stock Adjustment

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

Consider a retail firm entering any period with inventories, $y > 0$, of a homogenous good available for sale. The firm faces stochastic demand, ξ, drawn from a time-invariant distribution $F(\xi)$, and the value of its sales is p min $\{y, \xi\}$. At the end of the period, it may place an order $x > 0$ to increase its available stock for the next period; $y' = y - \min\{y, \xi\} + x$. The cost of any such order is $\phi + cx$, where $c > 0$ represents the unit cost of the good held in inventory. By proving K-concavity of the value function, Scarf was able to establish that the firm's optimal decision rule takes the following one-sided (*S*, *s*) form. (Scarf, 2005, shows this decision rule generalizes to a setting where the firm selectively sells its inventories with the option of leaving some demand unsatisfied. See Dixit, 1993, for a characterization of two- sided (*S, s*) policies arising in continuous time settings involving fixed and piecewise linear adjustment costs.)

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

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

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

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

Implications for Aggregate Investment

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

Caballero and Engel (1999) examine generalized (S, s) policies rationalized by stochastic fixed adjustment costs, ϕ, distributed i.i.d. across firms and over time. In this environment, a firm's capital, k, becomes part of its state vector alongside its total factor productivity, z. Moreover, microeconomic adjustment becomes probabilistic; firms with the same current gap between actual and target capital do not necessarily behave identically; rather, those with relatively low ϕ draws are more likely to alter their capital than those drawing high costs. If we transform

Caballero and Engel's gap-based analysis to reflect the firm-level state, (k, z), the implication is an adjustment hazard, $\Lambda(k, z)$, indicating what fraction of each group of firms sharing a common current state will choose to adjust their capital to a common target, $k^*(z)$. The resulting generalized (S, s) adjustment model allows convenient aggregation and has been studied in a variety of settings. (Dotsey et al. 1999, apply this basic framework to price adjustment, Thomas 2002, adopts it in an equilibrium business cycle model with lumpy investment, and King and Thomas 2006, use it to examine employment adjustment.)

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

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

More generally, this example illustrates that, when there is an aggregate shock, and thus a change in the target, higher moments of the distribution of capital across firms matter in determining movements in aggregate investment, because the adjustment hazard is a non-trivial function of capital. (This is an important distinction relative to the convex cost/partial adjustment model. Rotemberg (1987), shows its aggregate dynamics are reproduced by a model where individual firms adjust infrequently, but all face a common probability of undertaking adjustment, independent of their individual states. Given this constant hazard, only the first moment of the distribution is relevant in determining aggregate changes.) Alternatively, in the language of Caballero (1999, p. 841), microeconomic non-convexities can generate an important 'time-varying/history-dependent aggregate elasticity' of investment to shocks by allowing changes in the synchronization of firms' capital adjustments.

Although findings like those above echo throughout partial equilibrium studies involving lumpy adjustments, the omission of market-clearing relative prices (for example, equilibrium interest rates) may be critical to the inferred macroeconomic importance of non-convex factor adjustment costs. Significant aggregate nonlinearities can only occur if adjustment hazards exhibit large changes in response to shocks. Clearly, from the example above, such changes depend entirely on the extent to which $k^*(z)$ responds to changes in z. However, just as the capital adopted by a representative firm facing no adjustment costs varies far less when prices adjust to clear all markets, Thomas (2002) and Khan and Thomas (2003, 2006b) show that the target capital(s) selected by firms facing non-convex costs exhibit changes an order of magnitude smaller in general equilibrium. Because large movements in target capital, and hence in aggregate investment demand, would imply intolerable consumption volatility for households (at least in the closed-economy settings examined in these studies), they do not occur in equilibrium. Instead, small changes in relative prices serve to discourage sharp changes in $k^*(z)$, thereby preventing large synchronizations in firms' investment timing and leaving the aggregate series largely unaffected by the microeconomic lumpiness caused by non-convex adjustment costs.

Piecewise-Linear Costs

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

Partial irreversibilities have been widely examined in investment theory as an explanation for the common empirical finding that investment is insensitive to Tobin's q. Abel and Eberly (1994) characterize firm-level investment when the purchase price of capital, p_K^+, exceeds its sale price, p_K^- (and there are flow-fixed and convex adjustment costs). They show that this cost structure makes investment a nonlinear function of marginal q, implying a range of values over which the firm does not invest. (Veracierto 2002, solves a general equilibrium business cycle model where the resale price of capital goods is a constant fraction of the purchase price. Examining a wide range of values for this irreversibility parameter, he concludes that such frictions have no quantitatively significant effects for business cycle dynamics.) Elsewhere, in the context of employment adjustment, a simple example of piecewise-linear costs is an environment where firms incur no adjustment costs in increasing their employment, but pay a tax of $\phi > 0$ per worker fired. The implications of such firing costs for aggregate employment are theoretically ambiguous. While their direct effect is to discourage firing, they also induce a reluctance to hire. Bentolila and Bertola (1990) provide an early analysis suggesting that the direct effect dominates, while Hopenhayn and Rogerson (1993) find the converse.

Conclusion

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

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

See Also

- Inventory Investment
- Irreversible Investment

Bibliography

Abel, A. 1979. *Investment and the value of capital*. New York: Garland.

Abel, A., and J. Eberly. 1994. A unified model of investment under uncertainty. *American Economic Review* 84: 1369–1384.

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

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

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

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

Bertola, G., and R. Caballero. 1992. Irreversibility and aggregate investment. *Review of Economic Studies* 61: 223–246.

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

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

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

Caballero, R., and E. Engel. 1991. Dynamic (S,s) economies. *Econometrica* 59: 1659–1686.

Caballero, R., and E. Engel. 1993. Microeconomic adjustment hazard and aggregate dynamics. *Quarterly Journal of Economics* 82: 359–383.

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

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

Caballero, R., E. Engel, and J. Haltiwanger. 1997. Aggregate employment dynamics, building from microeconomic evidence. *American Economic Review* 87: 115–137.

Caplin, A. 1985. The variability of aggregate demand with (S,s) inventory policies. *Econometrica* 53: 1395–1410.

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

Caplin, A., and J. Leahy. 1997. Aggregation and optimization with state-dependent pricing. *Econometrica* 65: 601–626.

Caplin, A., and D. Spulber. 1987. Menu costs and the neutrality of money. *Quarterly Journal of Economics* 102: 703–725.

Chirinko, R. 1993. Business fixed investment spending: A critical survey of modelling strategies, empirical results, and policy implications. *Journal of Economic Literature* 31: 1875–1911.

Cooper, R., and J. Haltiwanger. 2006. On the nature of capital adjustment costs. *Review of Economic Studies* 73: 611–633.

Cooper, R., J. Haltiwanger, and L. Power. 1999. Machine replacement and the business cycle, lumps and bumps. *American Economic Review* 89: 921–946.

Dixit, A. 1993. Choosing among alternative discrete investment projects under uncertainty. *Economics Letters* 41: 265–268.

Dixit, A., and R. Pindyck. 1994. *Investment under uncertainty*. Princeton: Princeton University Press.

Doms, M., and T. Dunne. 1998. Capital adjustment patterns in manufacturing plants. *Review of Economic Dynamics* 1: 409–429.

Dotsey, M., R. King, and A. Wolman. 1999. State-dependent pricing and the general equilibrium dynamics of money and output. *Quarterly Journal of Economics* 114: 655–690.

Eisner, R., and R. Strotz. 1963. Determinants of business investment. In *Impacts of monetary policy*, ed. Commission on Money and Credit. Englewood Cliffs: Prentice–Hall.

Fisher, J., and A. Hornstein. 2000. (*S*, *s*) inventory policies in general equilibrium. *Review of Economic Studies* 67: 117–145.

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

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

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

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

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

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

Hamermesh, D. 1989. Labor demand and the structure of adjustment costs. *American Economic Review* 79: 674–689.

Hamermesh, D., and G. Pfann. 1996. Adjustment costs in factor demand. *Journal of Economic Literature* 34: 1264–1292.

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

Hayashi, F. 1982. Tobin's marginal q and average q: A neoclassical interpretation. *Econometrica* 50: 213–224.

Hopenhayn, H., and R. Rogerson. 1993. Job turnover and policy evaluation: A general equilibrium analysis. *Journal of Political Economy* 101: 915–938.

Jorgensen, D. 1963. Capital theory and investment behavior. *American Economic Review* 53: 247–257.

Khan, A., and J. Thomas. 2003. Nonconvex factor adjustments in equilibrium business cycle models: Do nonlinearities matter? *Journal of Monetary Economics* 50: 331–360.

Khan, A., and J. Thomas. 2006a. Inventories and the business cycle: An equilibrium analysis of (S,s) policies. Working paper.
Khan, A., and J. Thomas. 2006b. Idiosyncratic shocks and the role of nonconvexities in plant and aggregate investment dynamics. Working paper.
King, R., and J. Thomas. 2006. Partial adjustment without apology. *International Economic Review* 47: 779–809.
Koyck, L. 1954. *Distributed lags and investment analysis*. Amsterdam: North-Holland.
Lucas, R. 1967a. Adjustment costs and the theory of supply. *Journal of Political Economy* 75: 321–334.
Lucas, R. 1967b. Optimal investment policy and the flexible accelerator. *International Economic Review* 8: 78–85.
Lucas, R., and E. Prescott. 1971. Investment under uncertainty. *Econometrica* 39: 659–681.
Mendoza, E. 1991. Real business cycles in a small open economy. *American Economic Review* 81: 797–818.
Mosser, P.C. 1991. Trade inventories and (S,s). *Quarterly Journal of Economics* 106: 1267–1286.
Pindyck, R. 1988. Irreversible investment, capacity choice, and the value of the firm. *American Economic Review* 78: 969–985.
Ramey, V.A., and K.D. West. 1999. Inventories. In *Handbook of macroeconomics*, ed. J.B. Taylor and M. Woodford, vol. 1B. Amsterdam: North-Holland.
Rotemberg, J. 1987. The new Keynesian microfoundations. In *NBER macroeconomics annual 1987*, ed. S. Fischer. Cambridge, MA: MIT Press.
Sargent, T.J. 1978. Estimation of dynamic labor demand schedules under rational expectations. *Journal of Political Economy* 86: 1009–1044.
Scarf, H. 1960. The optimality of (S,s) policies in the dynamic inventory problem. In *Mathematical methods in the social sciences*, ed. K.J. Arrow, S. Karlin, and H. Scarf. Stanford: Stanford University Press.
Scarf, H. 2005. Optimal inventory policies when sales are discretionary. *International Journal of Production Economics* 93–94 (Special Issue): 111–119.
Sheshinski, E., and Y. Weiss. 1977. Inflation and costs of price adjustment. *Review of Economic Studies* 54: 287–303.
Sheshinski, E., and Y. Weiss. 1983. Optimum pricing policy under stochastic inflation. *Review of Economic Studies* 50: 513–529.
Thomas, J. 2002. Is lumpy investment relevant for the business cycle? *Journal of Political Economy* 110: 508–534.
Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit and Banking* 1: 15–29.
Treadway, A. 1971. The rational multivariate flexible accelerator. *Econometrica* 39: 845–855.
Veracierto, M. 2002. Plant-level irreversible investment and equilibrium business cycles. *American Economic Review* 92: 181–197.

Adjustment Processes and Stability

Franklin M. Fisher

Economic theory is pre-eminently a matter of equilibrium analysis. In particular, the centrepiece of the subject – general equilibrium theory – deals with the existence and efficiency properties of competitive equilibrium. Nor is this only an abstract matter. The principal policy insight of economics – that a competitive price system produces desirable results and that government interference will generally lead to an inefficient allocation of resources – rests on the intimate connections between competitive equilibrium and Pareto efficiency.

Yet the very power and elegance of equilibrium analysis often obscures the fact that it rests on a very uncertain foundation. We have no similarly elegant theory of what happens *out* of equilibrium, of how agents behave when their plans are frustrated. As a result, we have no rigorous basis for believing that equilibria can be achieved or maintained if disturbed. Unless one robs words of their meaning and defines every state of the world as an 'equilibrium' in the sense that agents do what they do instead of doing something else, there is no disguising the fact that this is a major lacuna in economic analysis.

Nor is that lacuna only important in microeconomics. For example, the Keynesian question of whether an economy can become trapped in a situation of underemployment is not merely a question of whether underemployment equilibria exist. It is also a question of whether such equilibria are stable. As such, its answer depends on the properties of the general (dis)equilibrium system which macroeconomic analysis attempts to summarize. Not surprisingly, modern attempts to deal with such systems have been increasingly forced to treat such familiar macroeconomic issues as the role of money.

We do, of course, have some idea as to how disequilibrium adjustment takes place. From Adam Smith's discussion of the 'Invisible Hand'

to the standard elementary textbook's treatment of the 'Law of Supply and Demand', economists have stressed how the perception of profit opportunities leads agents to act. What remains unclear is whether (as most economists believe) the pursuit of such profit opportunities in fact leads to equilibrium – more particularly, to a competitive equilibrium where such opportunities no longer exist. If one thinks of a competitive economy as a dynamic system driven by the self-seeking actions of individual agents, does that system have competitive equilibria as stable rest points? If so, are such equilibria attained so quickly that the system can be studied without attention to its disequilibrium behaviour? The answers to these crucial questions remain unclear.

A primary reason for that lack of clarity is the lack of a satisfactory theory about the disequilibrium behaviour of agents. A central example of the problem can be stated as follows. In perfect competition, all agents take prices as given. Then how can prices ever change? In a single market, for example, every firm believes that it will lose all its customers if it raises its price. Then who decides to go first when demand or cost increases? We are certain that such decisions are taken, but, at the level of satisfactory formal analysis, we do not know how.

While these issues arise in partial as well as general models, most of the literature on adjustment and stability has been at the general equilibrium level. (Search theory can be considered a partial modern-day exception.) Not surprisingly, that literature has largely begged the price-adjustment question, simply assuming that price somehow changes in the direction suggested by excess demand: $dp_i/dt = H^i(Z_i(p))$, where p is the vector of prices, $Z_i(p)$, the excess demand for the ith commodity, and $H^i(\cdot)$ a sign-preserving continuous function.

The question of who adjusts prices in this way is typically left unanswered or put aside with a reference to a fictitious Walrasian 'auctioneer'. That character does not appear in Walras (who did have prices adjusting to excess demands) but may have been invented by Schumpeter in lectures and introduced into the literature by Samuelson (who certainly did introduce the mathematical statement of price adjustment just given). Interestingly, however, the need for some such construct can reasonably be said to originate with Edgeworth, who wrote:

> You might suppose each dealer to write down his demand, how much of an article he would take at each price, without attempting to conceal his requirements; and these data having been furnished to a sort of market-machine, the price to be passionlessly evaluated. (1881, p. 30)

There has been only moderate progress since Edgeworth's day in explaining just what one is to suppose in considering anonymous price adjustment in competitive markets.

General equilibrium theory has taken its most analytically satisfactory form in the Arrow-Debreu world where all markets for present and future commodities open and close before any other economic activity actually takes place. Despite the lack of realism, this made it natural to consider adjustment processes in which only prices move (in the way described above) and trade, production, and consumption only occur after equilibrium is reached. Such a dynamic process is called 'tâtonnement', and the study of tâtonnement models dominated the stability literature until 1960. In that year, the publication of Herbert Scarf's counterexample (Scarf 1960) put an end to the hope that such models would turn out generally stable given only the ordinary assumptions of microeconomics. Tâtonnement stability requires extremely strong special assumptions.

This has extremely important implications. Indeed, it is not too strong to say that the entire theory of value is at stake. If stability requires trading (or production and consumption) to take place before equilibrium is reached, then the adjustment process itself changes the givens of the equilibrium problem (the endowments of agents, for example). This makes the set of equilibria also change in the course of adjustment, so that the equilibrium finally reached (assuming stability) differs from that computed by algorithms taking the initial situation as fixed. Moreover, comparative static analysis, that major tool of theory, will miscompute the effects of a displacement of equilibrium, for the equilibrium reached will depend on the adjustment process

and not merely on the displacement itself. While such effects may be small, they are certainly not known to be small. The argument that they are likely to be negligible because prices adjust much faster than quantities is unconvincing. The limiting case of such relative speeds of adjustment is tâtonnement and is known to lack general convergence properties.

The failure of tâtonnement was by no means the end of the stability literature, however. The early 1960s were marked by two important insights. These were: first, that considerable gains might be achieved by restricting the adjustment process itself rather than the excess demand functions of agents (Hahn 1961); second, that consideration of how trade takes place might lead to sensible restrictions. While logically separate, these two insights developed together in the study of 'non-tâtonnement processes', which are better called 'trading processes'.

In a pure-exchange trading process, prices continue to adjust as indicated by excess demands, but trade also takes place (consumption, however, still being postponed to equilibrium). The crucial question is how such trades should naturally be restricted, and here there are two leading candidates.

The first of these is the 'Edgeworth Process' (Uzawa 1962; Hahn 1962). Here the basic assumption is that trade takes place if and only if there is a group of agents, all of whom can gain in utility by trading among themselves at the current prices. With some complications due to the possibility that no such trades can be made at the initial configuration of prices, this assumption can be shown to generate a stable adjustment process. The crucial feature of the proof is that the sum of the utilities that would be achieved if trade ceases is increasing out of equilibrium, making that sum suitable for use as a Lyapunov function.

The basic assumption of the Edgeworth Process certainly seems attractive. Trade takes place because the agents participating make themselves better off thereby. Unfortunately, such attractiveness is somewhat superficial. First, the assumption places very large information requirements on the system. It is easy to construct examples where the only Pareto-improving trades require the participation of vast numbers of agents. While, as in the case of coalition formation in the theory of the core, the number of agents required cannot exceed the number of commodities, this is not a helpful limit when all future commodities are being traded. The assumption that trade readily takes place in such circumstances is not an easy one.

Second (and perhaps more important), the assumption that trade only takes place when participants each immediately gain in utility is only attractive when agents are supposed stupidly to expect prices constant and transactions to be completed. Once agents are allowed to become conscious of disequilibrium, transactions need not bring immediate utility gain; some transactions will be undertaken for speculative purposes, in the hopes that later transactions at profitable prices will materialize. While no rational agent ever trades without expecting to gain thereby, the basic assumption of the Edgeworth Process requires that every leg of a transaction bring a utility gain. It is crucial that the sum of the utilities agents would receive if trading ceased should always be increasing out of equilibrium. This is not true when arbitrage is involved – particularly when trade takes place for money. It is an open question whether the Edgeworth Process models can be adapted to allow more interesting behaviour.

The second major trading process model is the 'Hahn Process' (Hahn and Negishi 1962). Its basic assumption (sometimes known as the 'Orderly Markets Assumption') is as follows. After trade, there may be unsatisfied demanders of a particular commodity, say apples, or there may be unsatisfied suppliers of that commodity, but if markets are sufficiently well organized there will not be both. The Hahn Process assumes that potential apple buyers and potential apple sellers can find each other. Indeed, it might be said that this is what we mean when we speak of such buyers and sellers as being in the same 'market'. As a result, we assume that – after trade – any agent with a non-zero excess demand for some commodity finds that his or her excess demand for that commodity is of the same sign as that commodity's aggregate excess demand.

This has a powerful consequence. Since prices move in the same direction as aggregate excess

demand, any agent who cannot complete all planned transactions finds that the goods he or she would like to sell are falling in price, while the goods he or she would like to buy are becoming more expensive. The agent's target utility – the utility he or she would achieve if all transactions could be completed – is falling. As a result, out of equilibrium the sum of all target utilities falls and so can serve as a Lyapunov function. In effect, agents begin with unrealistically optimistic expectations and revise them downward until equilibrium is reached and expectations become mutually compatible. With some additional, relatively minor assumptions, the Hahn Process can be shown to be globally stable.

In fact, things are not so simple, for the assumption that buyers and sellers can find each other does not guarantee that unsatisfied excess demands for a given commodity will all have the same sign. This is because of the possibility that buyers will have nothing to offer that sellers are willing to accept. This problem cries out for the introduction of money as a medium of exchange (cf. Clower 1965). That introduction was accomplished by Arrow and Hahn (1971) who assumed that offers to buy must be backed up with money in order to be active and that prices are affected only by active, rather than target excess demands. Applying the Hahn Process assumption to active excess demands, the same global stability results can be obtained – provided one assumes that agents never run out of money. This 'Positive Cash Assumption' is very difficult to justify from more primitive ones in the context of naïve expectations.

The introduction of money raises other problems. In particular, unless money is included in the utility function, it is hard to see why agents plan to hold it in equilibrium. Nevertheless, such introduction is essential, particularly if firms are to be included. Without a common medium of exchange in which profits are measured, firms producing an oversupply of some good, say toothpaste, will have no incentive to sell it, reckoning profits in toothpaste rather than in dollars.

With money, however, the inclusion of firms in the Hahn Process model is fairly easy (Fisher 1974). Firms are assumed to sell promises to deliver outputs and acquire contracts to supply inputs, acting so as to maximize profits subject to ultimate production being feasible. Production itself is postponed until equilibrium (as is consumption in the pure exchange version). Again assuming that no household or firm ever runs out of cash, the target profits of firms decline if they cannot complete their planned transactions. Given that, the target utilities of the firms' owners – the households – also decline, and the stability result goes through much as before.

Despite its elegance, this is not a truly satisfactory result if one is interested in justifying the use of equilibrium economics. Apart from other difficulties, the equilibrium reached is one in which all trading opportunities have been exhausted. This is the consequence of working in an Arrow-Debreu framework, but it is not very satisfactory, and remains so even when some attempt is made to introduce production and consumption out of equilibrium (Fisher 1976). One would rather expect equilibrium to involve the carrying out of planned trades at correctly foreseen prices.

Further, the agents in trading-process models are remarkably stupid, always expecting prices to remain constant and transactions to be completed, when their constant experience tells them that this is not so. A model that hopes to explain how arbitraging agents drive a competitive economy to equilibrium can hardly afford to assume that agents do not perceive the very arbitrage opportunities that characterize disequilibrium.

An ambitious, though not altogether successful attempt to deal with these problems was made in the disequilibrium model of Fisher (1983). Agents have point expectations and are allowed to expect price changes. They take advantage of arbitrage opportunities, limited only by rules as to short sales and credit availability. Households maximize utility and firms profits, planning and engaging in consumption and production, respectively, in real time. Trade in firms' shares takes place both because of differing price expectations and because households purchase expected dividend streams as a way of transferring liquidity across time periods.

Agents also realize that they are restricted as to the size of their transactions. They make price

offers to get around such constraints. Thus each seller believes he or she faces a declining demand curve and has some monopoly power (similarly for buyers). The question of whether such perceptions disappear in equilibrium is the question of whether the equilibrium is Walrasian. In one form, it is also the question of whether there is equilibrium underemployment of resources. The answer turns out to be closely related to the extent to which the liquidity constraints are binding in equilibrium.

As this suggests, money plays a central role. The transactions demand for money does not disappear in equilibrium, which now involves the carrying out of previously planned transactions at the expected prices. On the other hand, 'money' in this model consists of very short-term bonds, bearing the same interest as all other assets in equilibrium. There is still no satisfactory theory in which agents hold non-interest-bearing bank notes in equilibrium.

Once one leaves equilibrium and leaves the theory of how the individual agent plans, matters become less satisfactory. This is largely because one has to deal with the behaviour of agents whose expectations are disappointed. The model handles this issue by making an extremely strong assumption called 'No Favourable Surprise'. This states that new, unexpected, favourable opportunities cease appearing. In effect, the kinds of shocks emphasized by Schumpeter (1911) – discovery of new products or processes, new ways of marketing, new sources of raw materials, and so forth – are ruled out if they are totally unforeseen. As in the Hahn Process, agents find that unexpected change makes them worse off as old opportunities disappear. With some technical complications, this ensures convergence to some equilibrium, although that equilibrium need not be Walrasian.

The problem is that 'No Favourable Surprise' is not a primitive assumption. One cannot hope to prove stability in a world constantly bombarded with exogenous Schumpeterian shocks. 'No Favourable Surprise', however, rules out the appearance of any unexpected opportunities, even those which arise in the course of adjustment to previous exogenous shocks. The Hahn Process model is a special case of this. So is the assumption of rational expectations. In a model with point expectations, however, rational expectations amounts to perfect foresight, and this begs the question of disequilibrium adjustment. It is unclear what happens under uncertainty and also unclear whether 'No Favourable Surprise' can be derived from other underlying premises.

Further, the very generality of the 'No Favourable Surprise' stability result has both satisfactory and unsatisfactory aspects. On the one hand, the price-adjustment mechanism left over from tâtonnement days can be dispensed with and individuals allowed to make price offers. On the other hand, just how those offers get made (or accepted) remains a mystery within the general confines of the 'No Favourable Surprise' assumption. We know that this depends on developing perceptions of demand and supply curves – of individual monopoly or monopsony power – but we do not know how those perceptions develop. As a consequence, the stability results give little insight into whether the system approaches a Walrasian or a non-Walrasian, quantity-constrained equilibrium. Similarly, we do not know the extent to which the adjustment process shifts the ultimate equilibrium or anything about adjustment speeds.

These remain questions of crucial importance for the under-pinnings of equilibrium analysis and, possibly, for the study of actual economies. They will remain unanswered without detailed analysis of how disequilibrium adjustment takes place when plans are frustrated. Equilibrium techniques will not succeed here, and new modes of analysis are needed if equilibrium economic theory is to have a satisfactory foundation.

See Also

- ▶ Auctioneer
- ▶ Economic Surplus and the Equimarginal Principle
- ▶ General Equilibrium
- ▶ Lyapunov Functions
- ▶ Stability
- ▶ Tâtonnement and Recontracting

References

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco/Edinburgh: Holden-Day/ Oliver & Boyd.

Clower, R.W. 1965. The Keynesian counterrevolution: A theoretical appraisal. In *The theory of interest rates*, ed. F.H. Hahn and F.P.R. Brechling. London/ New York: Macmillan/St. Martin's Press.

Edgeworth, F.Y. 1881. *Mathematical psychics,* Reprinted. New York: Augustus M. Kelley, 1967.

Fisher, F.M. 1974. The Hahn process with firms but no production. *Econometrica* 42: 471–486.

Fisher, F.M. 1976. A non-tâtonnement model with production and consumption. *Econometrica* 44: 907–938.

Fisher, F.M. 1983. *Disequilibrium foundations of equilibrium economics*. Cambridge: Cambridge University Press.

Hahn, F.H. 1961. A stable adjustment process for a competitive economy. *Review of Economic Studies* 29: 62–65.

Hahn, F.H. 1962. On the stability of pure exchange equilibrium. *International Economic Review* 3: 206–214.

Hahn, F.H., and T. Negishi. 1962. A theorem on non-tâtonnement stability. *Econometrica* 30: 463–469.

Scarf, H. 1960. Some examples of global instability of the competitive equilibrium. *International Economic Review* 1: 157–172.

Schumpeter, J. 1911. *The theory of economic development*. 4th printing of English trans. Cambridge, MA: Harvard University Press, 1951.

Uzawa, H. 1962. On the stability of Edgeworth's barter process. *International Economic Review* 3: 218–232.

Administered Prices

P. J. D. Wiles

Administered prices are prices set by enterprises, private or public, large or small, of their own volition in free markets for a period that they determine; so that prices do not fluctuate in the 'short run' with supply and demand. The market is cleared from moment to moment within this period by stock movements in the product and/or by queues of customers; and often by changes in production volume. The 'short run' includes periods long enough for it to *seem* bureaucratically possible to vary the price and thus to make more profit. Therefore *s.r.m.c./s.r.m.r.*, and administered prices constitute a failure to maximize profits. The administrator in this context is always the seller.

(A) We begin with the one great obvious exception: no prices in organized perfect markets, 'oriental' bazaars and auctions are administered. These markets do, of course, empirically exist: it is a myth that they are a myth. The principal examples are crops and metals (wholesale only); bonds, stocks and shares; foreign currency (wholesale only), houses; businesses (the latter two being disposed of in one-to-one higgling, as in an 'oriental bazaar'); most secondhand goods and antiques. Many homogeneous commodities, apt for organized perfect competition, are under state control; and many are dominated by oligopolies. But in perfect oligopoly, prices are not administered (the London Metal Exchange, in part).

Thus the textbook neoclassical description of the enterprise's price and output policy appears to be falsified by the very great majority of all turnover measured by value, and by the quite overwhelming majority of all individual prices. This has worried theorists far too little and they have not bothered to say much. But we must leave aside the reasons for that as belonging to methodological articles, and look at what might be said in favour of the text book. So here are the arguments that nevertheless long-run profits are being maximized despite, nay, because of, sticky prices (personal comments in brackets):

(B) (i) Very flexible prices annoy customers in an imperfectly competitive market; i.e. they lose good-will. So the firm that sets a not-too-high administered price gains customers and emerges with more total profit than it could gain from the average of the very short-run maximizing prices it could charge in an 'oriental' bazaar. (The writer knows of no research on this question, but finds the assertion plausible. It is alleged to be the foundation of many Quaker fortunes in the 17th century).

(B) (ii) In most businesses there are very many products indeed, and it is junior employees that charge prices for them. They cannot be trusted to higgle on behalf of the firm, and senior employees are too busy. So the latter provide a list or other document, which their juniors apply. This document simply cannot be altered every moment, so the administered price is the best that can be done towards maximizing profits. (This is correct; note that single-or-few-product enterprises do tend to belong to perfect markets).

And here (C) are the arguments that the phenomenon as defined does not exist:

(C) (i) Discounts from list price are extremely common. They are made ad hoc and *ad personam*, secretly. Therefore there are in fact no administered prices (the writer holds this to be a massive exaggeration of an admitted truth; but knows of no research that could tell us *how* massive, numerically speaking).

(C) (ii) Quality too can easily be manipulated in imperfect markets (but hardly – we must reply – in the short run, which is precisely how long an administered price is administered for).

Finally (D), here are the arguments for accepting that such prices are indeed a massive disproof of the doctrine that *m.c.* = *m.r.* even in the long run:

(D) (i) There is little empirical evidence that entrepreneurs or managers *ever* maximize even their long-run profits. The basic proposition of all neoclassical economics has never been properly researched; it has simply been elevated (or degraded?) into an axiom. We should reject its high philosophical claim and simply use our eyes. Our eyes may indeed confirm it, but only as an empirical generalization.

(D) (ii) Nevertheless it does seem probable, on mere inspection, that entrepreneurs and managers maximize their short- and simultaneously their long-run profits in perfect competition: what else could one be doing in a market where one is concerned with few products (see B (ii)), and there is no good-will (see B (i))? However that leaves wholly intact the possibility that entrepreneurs and managers behave differently in other types of market.

(D) (iii) In psychological terms homo economicus is a psychopath, though in situations (A) he has little choice. Now admittedly psychopathy is an arbitrary term, to be used with extreme care. But every psychologist would pale before calling almost all men psychopaths. There is nothing whatever in the other social sciences to indicate that profit-maximization is in fact a human norm.

(D) (iv) So far we have merely cleared the ground. The first positive argument is that it is obvious that when, in situations other than (A), we are making losses we do maximize our profits. But the fact that prices become more flexible in depression *confirms* that both long- and short-run profit-maximization are, in the majority of market situations, optional; for it entails that when there is no depression they become less flexible again. The often urged greater survival value of profit maximization refers to survival circumstances, not all circumstances.

(D) (v) Then there is the undoubted fact that FIFO accountancy influences prices. Ordinary observation tells us that in a period of prolonged but not very rapid price-rise the goods on sale in a shop are all priced by applying the customary gross profit margin (absolute or relative) to the historical cost of acquisition of the particular physical batch; so that on one shelf or in one drawer the identical object has different prices. It is very difficult indeed to attribute this to administrative difficulty – why not simply put a single general price label on the shelf or drawer, leaving

individual items unmarked? And it certainly is not profit maximization. Again the business pages of the newspapers consistently refer to 'cost increases coming up through the pipe-line'; this banal, and generally accepted, phrase means the same thing.

(D) (vi) The happy hunting grounds of administered prices are manufacturing, retailing and the standard services of transportation. In construction 'cost-plus', where profit is an agreed percentage of whatever cost will turn out to be, is very notoriously the main method of price formation; still more so in pricing modifications to contracts. But cost-plus is the essence of practical price administration. R&D projects are also priced on 'cost-plus'; so indeed are all prototype machines and all non-standard repair jobs. 'Cost-plus' arises out of uncertainty as to costs; it induces profit-maximizers to raise their costs above the minimum for the contracted output. But to an ordinary customer like the writer it is evident that a high proportion of cost-plus chargers do not abuse their position.

In some situations *l.r.p.m.* demands *s.r.p.m.* These are the totally impersonal, or at least one-off, market situations listed above (A). In these, mere inspection, as we saw, tells us that *s.r.p.m.* is practised; and it follows that *l.r.p.m.* is too, since there is no good-will to be lost. But in situations (B) *l.r.p.m.* forbids *s.r.p.m.*, since it loses good-will, and may also be too great an administrative burden. Now while again mere inspection tells us that *s.r.* profit is not being maximized, we may not infer that *l.r.p.m.* is being maximized. That is not evident, but can only be proved (or disproved) empirically.

Is satisficing a failure to maximize long-run profit? Certainly satisficing is implicit at every point above where administered prices are described. But if we accept it wholeheartedly it tells us nothing more. It means that we *pay attention to* the various costs of search: not that we minimize them (or indeed any other costs); not that we maximize profit net of search costs, nor again that we don't.

The trouble with satisficing is that it is all things to all men. Profit maximizing is an ex ante, or policy, concept; it requires, in the legal phrase, *mens rea* and cannot easily be proved or disproved from observation. This is truer the longer the run we consider and the more space we leave for human judgement. But satisficing is by itself simply a technique not a policy: a recognition that the setting of qualities, prices and outputs requires serious research and thought, but that a decision must come soon. Thus in terms of 'soonness' we have many courses open to us. If we decide very soon, we may be consciously following the full cost principle, and merely avoiding losses; or consciously maximizing profit but making a mistake about the time and resources required for optimization; or some third thing. If we decide at the 'right' moment, we may be consciously maximizing a concept of profit that includes decision-making costs and benefits; or consciously applying full cost but dithering too long; or again some third thing. There can also be systematic error as to how to maximize profits, despite a genuine wish to do it – notably ignorance of the marginal analysis at entrepreneurial level. But that undoubted fact is not quite what here concerns us. For those who do understand it still administer their prices. It should be remembered that a mere loss-avoider must also satisfice. He too faces an intellectual and information problem, though a simpler one: he too must eventually cut off his research and decide, though that point comes sooner. The sales-maximizer-subject-to-minimum-return (à la Baumol) is in the same boat too: with a problem of intermediate complexity. Satisficing, to repeat, is a universal tool.

Does *oligopoly* account for the whole phenomenon? Clearly not perfect oligopoly, but many have been tempted by the kinked demand curve of imperfect oligopoly, which gives such latitude to a price-setter. Hall and Hitch (1939), the pioneers, certainly rested their work on this, and so does Sylos-Labini (1979 passim). But in reality the phenomenon is much more widespread, because the demand and marginal cost curves are *uncertain* also in monopolistic competition

and monopoly, where kinks are unknown. Indeed the curves, though continuous, are thick bands and not narrow lines at all. This has the same effect as the kink, though for slightly different reasons. For the kink forms wherever the price happens to be, and the price may have been rationally set originally; but where the narrow lines become thick bands the situation is indeterminate a priori.

Nothing, then, except a methodologically false tradition forbids us to say that normal price-setting is merely cost-covering, or loss-avoidance plus a decent allowance for net profit; and that the quantification of the word 'decent' is a purely empirical task. It may, for all we know, differ more according to the ideology, nationality, religion or historical epoch of the entrepreneur/manager than according to the state (depressed or active) or the form (imperfect oligopoly, monopoly etc.) of the market. The work has not been done: we do not know.

Now the price that yields a 'decent' profit may be either lower or higher than the profit-maximizing price. It must however lie *below* the latter price much the most often. Why incur obloquy when for the same money you can be popular? The notion that prices administered in a long-run-non-profit-maximizing spirit are usually 'too' low is used by Baumol (1969, pp. 47–52, 63–6) to explain his observation that firms prefer sales volume to profit volume. In so doing he distances himself, to be sure, from the original Oxford full cost doctrine, but not importantly.

The 'lowness' of prices explains also cost inflation. In a climate of general price rises, where this is the general expectation for a long time ahead, the 'decent' price has a consequence quite incompatible with long-run profit maximization. The perpetual small cost increases, due to rises in import prices, wage-rates, variable and even fixed taxes, domestic fuel and raw material prices etc., can be accommodated without a damaging output shrinkage simply by raising the output price *towards* its profit-maximizing level. This is what cost inflation is. It never occurs in type A markets.

This is evidence indeed for a sceptical attitude towards neoclassical microeconomics. But better evidence would be more candid and straightforward empirical research, that simply treated homo economicus as a hypothesis like any other. The type A market reminds us that the hypothesis could easily be confirmed on many occasions.

The full-cost principle in particular, and administered prices in general, seem to be methodologically offensive to orthodox economic theory. No facts should be that. The immunity-system of the neoclassical body rejects every attempted transplant. Why? First, the whole theory, or generalization, is crude, indeterminate, superficial and unintellectual. The full-cost principle, or the cost-plus determination of all non-perfect-market prices from the fairly competitive parts of the private sector right through without distinction to the most protectedly monopolistic parts of the public sector, is the theory of value of the man in the street, and of most people in authority who set prices.

Yet, secondly, those four epithets above do not signify falsity. Many a good economist slips into such language *obiter*. The following passage from Okun, before he changed sides (1960, pp. 35–6), picked virtually at random from all the literature of economics, shows the same unthinking reflex:

> The main element in the stubborn climb of prices and wages through most of 1969 was the enormous strength of demand for labour. After years of operating in a tight labour market, businessmen hired aggressively both to catch up and get ahead. They added far more workers to their payrolls than would have been dictated merely by short-run needs. Between mid-1968 and mid-1969, for example, wholesale and retail trade added 600,000 employees or a 4.5 per cent rise in their work force, while the volume of real goods flowing through trade barely increased. Such personnel policies get reflected in sagging productivity, a substantial addition to unit labour costs, and continued tightness in labour markets; the result is more inflationary pressure on both prices and wages.

The addition to demand through higher wages were certainly not upper-most in this author's mind. Yet the whole 53-page chapter (entitled Inflation: Problems and Prospects) is orthodoxly on the side of demand inflation, fiscal and monetary management, etc.

The idea has its place within the history of economic doctrine. It was dominant – unconsciously – in Smith, Ricardo and Marx. As pre-marginalist economists, they wrote always

as if it were true, though they really had no systematic micro-economics. The idea certainly animated Marshall, an early and undogmatic marginalist who seems to have believed in loss minimization only. (For these 'forefathers', cf. Wiles 1961).

In the 1930s marginalism was completed by 'imperfect competition'; the marginalization of revenue rounded off that of cost. This both proved the necessity of a theory of value related to the theory of the firm, and ensured that it should be tidy and determinate – so non-empirical. Therefore, it has been hostile to these ideas. So much is evident all the way from William Stanley Jevons to Joan Robinson. The realistic pre-war reaction to marginalism came very shortly after Robinson, who may be said to have caused it. The reaction was called the full cost principle in UK, administered prices in USA. The former analysis was more complete since it had of course administered prices (Hall and Hitch 1951 – originally 1938), but included a long analysis of costs. The latter was short in this latter respect, and so had too little regard for its own micro-foundations, being macro-economically biased, but made the contrast with perfect competition more clear (Means 1935; cf. Sylos-Labini 1969, p. 110). The erosion of both traditions after World War II is described in some detail by Lee 1984. Yet they survive in business departments (e.g. Jackson 1982), through not in 'industrial economics' courses within departments of economics. There, the view became widespread that all prices are somehow administered so as to maximize long-run profit, and full cost is an awkward and uneducated *language* only, for describing essentially marginalist decisions (Lee, op. cit.). The much praised text of Koutsoyiannis (1979) must be placed here in the last analysis – daring and unusual as she was to include such a subject in a textbook at all. Strong exceptions are Okun (1981, ch. 4 and p. 223) and Baumol (1967, pp. 48–52).

See Also

- ▶ Cost-Push Inflation
- ▶ Kinked Demand Curve
- ▶ Satisficing

Bibliography

Baumol, W. 1966. *Business behavior, value and growth*, Revisedth ed. Princeton: Princeton University Press. 1967.

Hall, R.E., and C.J. Hitch. 1939. Price theory and business behaviour. In *Oxford studies in the price mechanism*, ed. P.W.S. Andrews and T. Wilson. Oxford: Oxford University Press. 1951.

Jackson, D. 1982. *Introduction to economics: Theory and data*. London: Macmillan.

Koutsoyiannis, A. 1979. *Modern micro-economics*, 2nd ed. London: Macmillan.

Lee, F. 1984. Whatever happened to the full-cost principle (USA). In *Economics in disarray*, ed. P.J.D. Wiles and G. Routh. Oxford: Blackwell.

Means, G.C. 1935. Industrial prices and their relative inflexibility. *Senate Document No. 13*. 74th Congress, 1st Session, Washington, DC.

Okun, A.M. 1970. Inflation, problems and prospects. In *Inflation: The problems it creates and the policies it requires*, ed. A.M. Okun, H.M. Fowler, and M. Gilbert. New York: New York University Press.

Okun, A.M. 1981. *Prices and quantities*. Oxford: Blackwell.

Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.

Sylos-Labini, P. 1969. *Oligopoly and technical progress*. Cambridge, MA: Harvard University Press.

Wiles, P.J.D. 1961. *Price, cost and output*, 2nd ed. Oxford: Blackwell.

Advances

G. Vaggi

The French Physiocrats used the term *avances* to indicate the outlays which had to be used in the process of production in order to yield a return in the future. In the *Essai sur la nature du commerce en général* (1755) Cantillon had already used the term advances, but it becomes prominent only in Physiocratic literature. The way in which Quesnay used this term clearly indicates that he was referring to what is now called either capital or means of production. The advances can be regarded as a sum of money, but more frequently the Physiocrats referred to the commodities which had to be 'advanced' in order to carry on the process of production. The different types of advances depend

upon the methods of production adopted, which establish the relationships between the inputs and the output in each sector of the economy.

The Physiocrats dedicated particular attention to agriculture, where they distinguish three types of advances. The ground advances, *avances foncières*, include all the expenses which are necessary in order to prepare the soil for cultivation: drainage, cleaning of the soil, transportation and housing facilities (see Baudeau 1767, pp. 154–6). This kind of advance had to be made once and for all; they were a sort of prerequisite for cultivation. Most of them were made by the landowners, whose rents were a compensation for this initial contribution to production (ibid.). The Physiocrats emphasize the importance of the advances of the agricultural entrepreneur, the farmer, which can be divided into two categories: the original advances, *avances primitives*, and the annual ones, *avances annuelles*. The former group includes all the instruments of cultivation like carts, ploughs etc., which can be used for many years and need annual repairs and maintenance. However, according to the Physiocrats the original advances also include the horses employed in cultivation and their fodder (see Quesnay 1758, p. vi; Meek 1962, p. 279). These original advances are like fixed capital, and they are assumed to wear out at a rate of 10 per cent a year. Thus, the farmers must use part of the annual output of cultivation to keep the stock of these advances at its initial level (see Quesnay 1766, p. 152). The annual advances of agriculture are the raw materials, seeds etc. and the consumption goods necessary to allow the peasants and their families to work until the next harvest. These commodities are entirely consumed during the process of production and as such they must be regarded as circulating capital. In order to maintain the level of agricultural activity unchanged it is necessary to replace the entire annual advances. The whole annual advances plus the interest required to preserve the original ones from wear and tear make up the annual returns of cultivation, the *réprises* (see Quesnay 1766, p. 154). The returns indicate which part of annual production must be set aside in order to be employed in the following production period. According to this way of examining the process of production, each year the annual output of agriculture must include all the types of commodities which have been used up during the previous productive process as advances. The Physiocratic concept of advances is then clearly linked to their view of the economy as a system which regularly reproduces itself. The part of the social product which is in excess of the returns is the surplus, or net product.

Quesnay used the concept of advances to establish some precise numerical relationships between the inputs and the output of the production process both in agriculture and in manufacturing. For instance, he believed that the best methods of cultivation required a ratio of one to five between the annual and original advances. A modern agricultural sector must have a large stock of original advances, which allows a net product equal to the amount of annual advances. Thus modern techniques of cultivation yield a revenue, or surplus, of 100 per cent (see Quesnay 1766, p. 151). According to Quesnay the industrial sector employed only annual advances and its output is exactly equal to the value of these advances, thus there is no surplus. The notion of advances is an important element of the Physiocratic doctrine that only agriculture yields a net product.

Turgot, too, employed the concept of advances, but he did not clearly distinguish the analysis of the advances of agriculture from those of the other secotrs of the economy (Turgot 1766, pp. 147, 151). Turgot adopts Quesnay's definitions of annual and original advances as those commodities which must exist at the beginning of the process of production (ibid., pp. 153–4). For Turgot the term 'advances' also refers to the employment of money in one of the several types of investments; he also uses the terms 'moveable wealth' and 'capital' instead of that of 'advances' (see ibid., pp. 145, 152).

Adam Smith substitutes the term 'capital' for 'advances', even though he still uses Quesnay's notion of means of production, which must be advanced by the entrepreneur in order to carry on the process of production. Smith distinguishes fixed and circulating capital. The former notion refers to all the machines and the instruments

which yield a profit to the entrepreneur without being sold. Circulating capital indicates all the commodities which yield a profit only when they are sold at the end of the productive process, but not when they still belong to the capitalist (see Smith 1776, 1976, vol. I, p. 279). However, contrary to Quesnay, in his theory of value Smith emphasizes the role of circulating capital. Thus the overall capital of society seems to be entirely made up by the wages advanced to the workers (ibid., pp. 66–7nh, pp. 110–11), because all machines are ultimately produced by labour.

Ricardo distinguishes fixed and circulating capital according to the durability of the input examined, but he admits that it is often difficult to draw a precise distinction between the two notions (Ricardo 1821, pp. 31–2, 150–1). Ricardo accepts Smith's idea that from the point of view of society as a whole capital is made up of the value of the wages advanced to productive workers (ibid.). Malthus was the last classical economist to use the term 'advances', by which he meant all the commodities which had been accumulated in the past and whose value had to be subtracted from that of annual production in order to measure the profits of the entrepreneurs.

See Also

► Physiocracy

References

Baudeau, N. 1767. Explication du Tableau économique à Madame de ***, par l'auteur des Ephémérides. In *Ephémérides du citoyen*, vols. XI and XII.

Malthus, T.R. 1820. Principles of political economy. In *The works and correspondence of David Ricardo*, vol. 2, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Meek, R.L. 1962. *The economics of physiocracy.* London: Allen & Unwin.

Quesnay, F. 1758. Explication du Tableau èconomique. In *Quesnay's Tableau Economique*, ed. M. Kuczynski and R. L. Meek. London: Macmillan, 1972.

Quesnay, F. 1766. Analyse de la formule arithmétique du Tableau économique. In Meek (1962).

Ricardo, D. 1821. In *On the principles of political economy and taxation*, 3rd ed, vol. 1. *The works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of Nations*. Oxford: Oxford University Press, 1976.

Turgot, A.R.J. 1766. Réflexions sur la formation et la distribution des richesses. In *Turgot on progress, sociology and economics*, ed. R.L. Meek. Cambridge: Cambridge University Press, 1973.

Adverse Selection

Charles Wilson

Abstract

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

Keywords

Adverse selection; Akerlof, G.; Asymmetric information; Bertrand game; Credit rationing; Incentive compatibility; Insurance markets; Pareto improvement; Rationing; Resale markets; Reservation price; Self-selection; Separating equilibrium; Signalling; Walrasian equilibria

JEL Classifications

D8

Adverse selection refers to a negative bias in the quality of goods or services offered for exchange when variations in the quality of individual goods

can be observed by only one side of the market. For instance, suppose sellers of high-quality goods have a higher reservation price than sellers of low-quality goods, but that buyers cannot directly determine the quality of a specific good offered for sale. Then any mix of goods offered for sale at the market price must include the low-quality goods. That is, the market adversely selects for low-quality products.

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

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

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

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

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

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

All of these sources of inefficiency can be illustrated with a slight modification to Akerlof's example. Suppose we change the distribution of

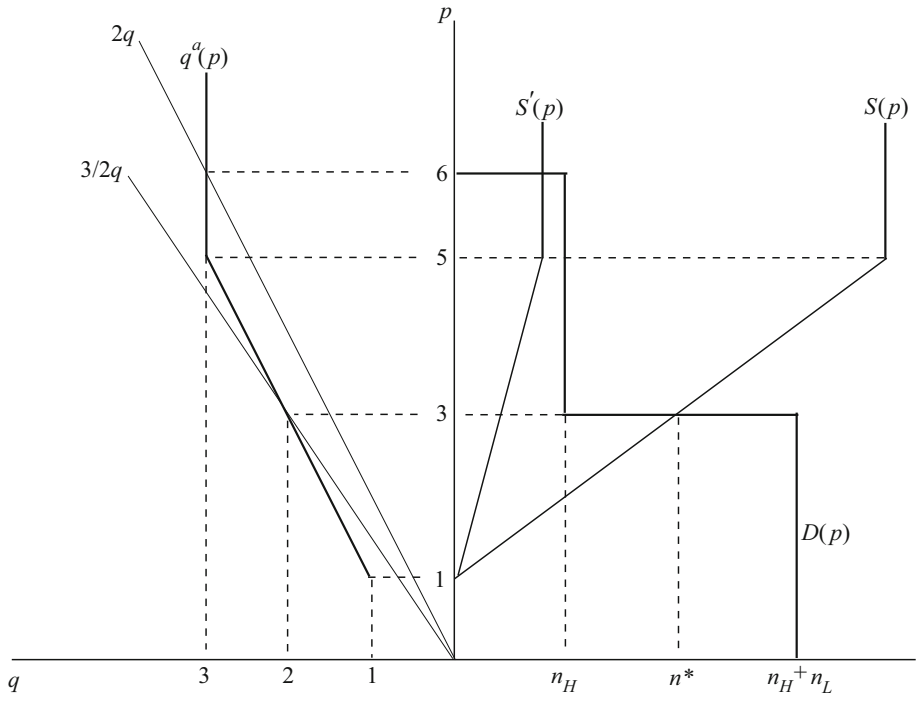
the n_s cars so that q is uniformly distributed between 1 and 5. Then, at any price p between 1 and 5, the supply of cars is $S(p) = \frac{p-1}{4}\, n_s$ and average quality is $q^a(p) = \frac{1+p}{2}$. At any price $p > 5$, $S(p) = n_s$ and $q^a(p) = 3$. On the demand side, we suppose there are two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

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

At prices 3 and 6, demand is a correspondence.

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

Observe that this allocation exhibits all of the sources of inefficiency that were identified above. First, not all potential buyers purchase a car even though half of the cars remain unsold, all of which are more valuable to buyers than to owners. Second, the cars that are sold provide the least possible net benefit to buyers. If only half of the cars are to be sold, efficiency requires they be the highest-quality cars. Third, since all buyers purchase from the same pool of cars, the cars that are sold are not efficiently allocated among buyers. Since high-intensity buyers value quality more than low-intensity buyers, the efficient allocation of these



Adverse Selection, Fig. 1 An inefficient Walrasian allocation

cars requires that the high-intensity buyers receive the cars with the highest quality.

Given the asymmetry in information, there is typically no incentive-compatible mechanism that achieves first-best efficiency. However, there may be instruments or mechanisms that may increase net surplus and in some cases even generate a Pareto improvement. For instance, for supply curve $S(p)$ a subsidy on sales would increase the volume of trade. However, the resulting allocation would not be completely efficient since low-quality cars are still sold before high-quality cars and both types of buyers still purchase from the same pool of cars. We explore below some other mechanisms that may be used to further improve efficiency.

Multiple Walrasian Equilibria

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

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

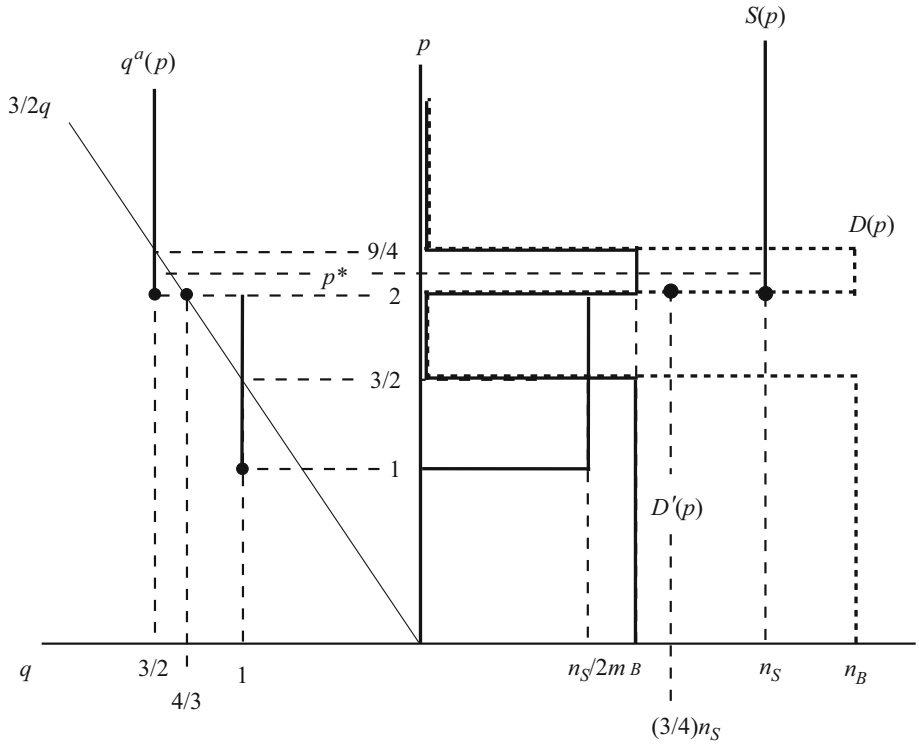
In this example, multiple Walrasian equilibria arise whenever the number of buyers exceeds the total number of cars. Such a case is illustrated in Fig. 2, where demand $D(p)$, indicated by the heavy dotted line, intersects $S(p)$ at prices $\frac{3}{2}$, 2, and $\frac{9}{4}$. All cars are sold at price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

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

Pareto Improving Price Floors

Because of the dependence of average quality on price, it is sometimes possible to achieve an additional Pareto improvement by setting a price floor and rationing the excess supply of cars. Consider again the example illustrated in Fig. 2. If we reduce the number of buyers to m_B where $\frac{n_S}{2} < m_B < n_S$, then we obtain a demand curve like $D'(p)$, illustrated by a heavy solid line. In this case, there is only one Walrasian equilibrium at price $p = \frac{3}{2}$. At this price, only low-quality cars are offered for sale and buyers gain no net benefit.

Now suppose that we impose a floor ceiling at some price p^* between 2 and $\frac{9}{4}$ .Since high-quality cars are also supplied at this price, average quality rises to $q^a(p^*) = \frac{3}{2}$ which provides any buyer with a positive net benefit. Since there are more sellers than buyers at this price, sales must be rationed. Nevertheless, owners of both low-quality and high-quality cars benefit from selling at this price. Owners of high-quality cars benefit because the Walrasian price is below their reservation value. And since more than half of the cars are sold at this price, the expected return to low-quality sellers is also higher at price p^*. At the Walrasian price $p = \frac{3}{2}$, their net benefit from a sale



Adverse Selection, Fig. 2 Multiple Walrasian equilibria

is $\frac{1}{2}$, while at the price floor $p^* > 2$, their net benefit from a sale is at least 1.

Uninformed Price Stters and Rationing

Our analysis so far has focused on primarily on Walrasian allocations. In a frictionless economy with perfect information and a large number of competing agents, this solution is generally robustly independent of the mechanism or conventions by which the price is set. However, once we introduce asymmetries in information, the opportunity for market participants to exploit the relation between quality and price or to indirectly identify products of different quality may lead to different market behaviour. To study these effects, we need to be more explicit in specifying the mechanism by which trade takes place.

Consider a market mechanism in which each buyer fixes a price at which he is willing to buy. To sell their cars, sellers first queue at the highest announced price. Any excess supply then spills over to successively lower-price offers until the supply of cars is exhausted or there are no more offers to buy. Buyers who announce a price below the point at which supply is exhausted do not obtain a car.

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

Figure 2 illustrate two types of solution to this problem. For the case where the number of buyers is $n_B > n_S$, represented by the heavy dotted demand curve $D(p)$, the equilibrium price is $p = \frac{9}{4}$,

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

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

Informed Price Setters

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

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

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

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

contracts to identify product quality. A market for insurance provides a good example.

Self-selection in Insurance Markets

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

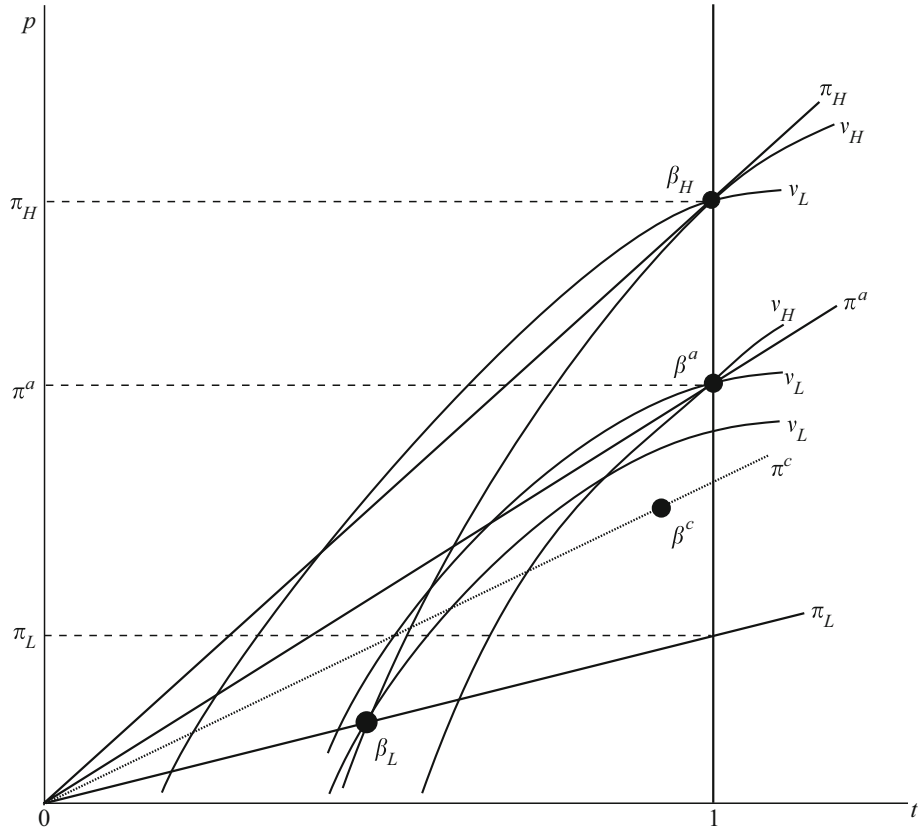
Suppose there are two types of insurance consumers. Each consumer has the same risk-averse von Neumann-Morgenstern utility u, the same initial wealth W and the same reduction in wealth to $W - 1$ in case of an accident. Low-risk types have an accident with probability π_L and high-risk types have an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

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

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

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

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

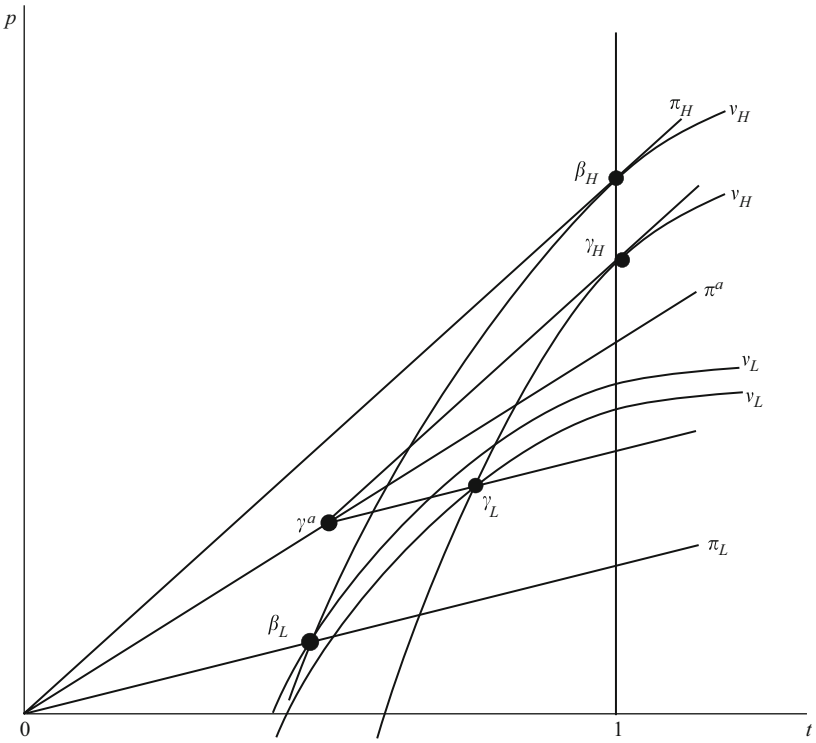


Adverse Selection, Fig. 3 Equilibrium in an insurance market

high-risk types and the best zero-profit policy for the low-risk types, subject to the self-selection constraint for high-risk types to choose policy β_H.

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

If the aggregate zero-profit line intersects the low-risk indifference curve passing through π_L, as illustrated by the dotted line labelled π^c in Fig. 3, then there is no equilibrium for this game. In this case, a firm may offer a policy just above β^c that attracts both types of consumers and still makes positive profits in the aggregate. If we permit individual firms to offer a menu of contracts as in Miyazaki (1977), then equilibrium fails to exist under an even wider range of parameters. A number of authors have suggested alternative solution concepts, incorporating non-Nash behaviour, that generate an equilibrium for this case. Wilson (1977) defines a solution concept in which both types purchase a policy like β^c. Riley (1975) proposes an alternative solution concept for which the separating allocation β_L and β_H is an equilibrium.



Adverse Selection, Fig. 4 The public provision of insurance

Efficient Public Provision of Insurance

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

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

See Also

- ▶ Credit Rationing
- ▶ Implicit Contracts
- ▶ Incomplete Contracts
- ▶ Moral Hazard
- ▶ Selection Bias and Self-Selection
- ▶ Signalling and Screening

Bibliography

Akerlof, G. 1970. The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84: 488–500.
Cho, I., and D. Kreps. 1987. Signalling games and stable equilibria. *Quarterly Journal of Economics* 102: 179–221.
Hendel, I., and A. Lizzeri. 1999. Adverse selection in durable goods markets. *American Economic Review* 89: 1097–1115.
Miyazaki, H. 1977. The rat race and internal labor markets. *Bell Journal of Economics* 8: 394–418.
Riley, J.G. 1975. Competitive signalling. *Journal of Economic Theory* 10: 175–186.
Rothschild, M., and J. Stiglitz. 1996. Equilibrium in competitive insurance markets: An essay on the economics imperfect information. *Quarterly Journal of Economics* 90: 629–649.
Spence, M. 1973. Job market signalling. *Quarterly Journal of Economics* 87: 355–374.
Stiglitz, J., and A. Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71: 393–410.
Wilson, C. 1977. A model of insurance with incomplete information. *Journal of Economic Theory* 16: 167–207.
Wilson, C. 1980. The nature of equilibrium in markets with adverse selection. *Bell Journal of Economics* 11: 108–130.

Advertising

Richard Schmalensee

Abstract

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

Keywords

Advertising; Concentration; Market structure; Entry

JEL Classifications

D00

Advertising has been controversial, probably more so that its economic importance would justify, at least since the emergence of the mass media in the 19th century. In the United States, advertising spending in the second half of the 20th century was just above two per cent of GDP. This ratio grew slowly over time; it is much lower in most other countries, especially in developing nations. In the United States and elsewhere, the ratio of advertising to sales varies dramatically among industries, even if attention is limited to industries selling consumer goods and services.

Chamberlin's Theory of Monopolistic Competition (Chamberlin 1933) was the first major work in economics to treat advertising formally, but its analysis led to few definite positive or normative conclusions. Perhaps reflecting the traditional distaste for advertising in the intellectual community, most early discussions of advertising by economists were generally critical, describing it as wasteful, manipulative, and anticompetitive. Its main redeeming feature was that it provided a source of revenue for the press (Kaldor 1950, is a leading example). Most writers are less enthusiastic about the relation between advertising and the media, perhaps because of the rise of television.

Consumer Demand

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

Following Nelson (1974), a number of authors have explored the possibility that advertising affects behaviour through such signals. The core of the argument is that advertising is more profitable for high-quality than low-quality producers, all else equal, since the former are more likely to

enjoy repeat sales. In sharp contrast, information processing models of human behaviour, explored in the marketing literature, suggest that advertising may affect behaviour mainly by enhancing a brand's chances of being on the short list ('evoked set') from which final choices are made.

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

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

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

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

Seller Behaviour

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

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

Consideration of these last two points indicates that the intensity of advertising may rise or fall with increases in market concentration (Schmalensee 1972, ch. 2). On the one hand,

reductions in the number of sellers would be expected to reduce the intensity of all forms of rivalry, and thus to reduce advertising spending. On the other hand, if sellers in concentrated markets manage to raise prices far above marginal costs, they thereby enhance incentives to advertise.

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

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

Economic Welfare

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

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

In practice, bureaucrats are far from omniscient, and the way in which information is presented to consumers affects the extent to which they retain and use it. Advertisers have every incentive to present information effectively, though they rarely have any incentive to present all information that might affect decisions. Advertising, like democracy, is terrible in principle but better than any known alternative in practice. Note also that advertising is practised, though not intensively by US standards, in socialist economies.

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

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

Market Structure

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

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

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

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

But profits are high when price–cost margins are large, and large margins encourage advertising (Schmalensee 1972, ch. 7). Since it is difficult to model advertising spending decisions empirically, it is difficult to deal adequately with this simultaneous equations problem. Moreover, accounting measures of profit treat advertising as an expense, but it should be treated as a durable investment if its effects on demand persist over time. If those effects are assumed to be very long- lived, correcting the accounting profitability figures eliminates the correlation with advertising. Unfortunately, like so much in this area, the longevity of the impact of advertising on demand remains controversial.

New Empirical Developments

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

See Also

- ▶ Chamberlin, Edward Hastings (1899–1967)
- ▶ Market Structure
- ▶ Monopolistic Competition

Bibliography

Benham, L. 1972. The effects of advertising on the price of eye-glasses. *Journal of Law and Economics* 15: 337–352.

Borden, N.H. 1942. *The economic effects of advertising*. Chicago: Irwin.

Butters, G. 1977. Equilibrium distribution of sales and advertising prices. *Review of Economic Studies* 44: 465–491.

Camanor, W.S., and T.A. Wilson. 1979. The effect of advertising on competition: A survey. *Journal of Economic Literature* 17: 453–476.

Chamberlin, E.H. 1933. *The theory of monopolistic competition*. Cambridge, MA: Harvard University Press.
Dixit, A., and V. Norman. 1978. Advertising and welfare. *Bell Journal of Economics* 9: 1–17.
Kaldor, N. 1950. The economic aspects of advertising. *Review of Economic Studies* 18: 1–27.
Lambin, J.J. 1976. *Advertising, competition, and market conduct in oligopoly over time*. Amsterdam: North-Holland.
Nelson, P. 1974. Advertising as information. *Journal of Political Economy* 82: 729–754.
Porter, M. 1976. *Interbrand choice, strategy, and bilateral market power*. Cambridge, MA: Harvard University Press.
Schmalensee, R. 1972. *The economics of advertising*. Amsterdam: North-Holland.
Sethi, S.P. 1977. Dynamic optimal control models of advertising. *SIAM Review* 19: 685–725.
Telser, L.G. 1964. Advertising and competition. *Journal of Political Economy* 72: 537–562.

Bibliographic Addendum

Ackerberg, D.A. 2001. Empirically distinguishing informative and prestige effects of advertising. *RAND Journal of Economics* 32: 316–333.
Benham, L. 1972. The effect of advertising and the price of eyeglasses. *Journal of Law and Economics* 15: 337–352.
Cady, J.F. 1976. An estimate of the price effects of restrictions on drug price advertising. *Economic Inquiry* 14: 493–510.
Devine, D.G., and B.W. Marion. 1979. The influence of consumer price information on retail pricing and consumer behavior. *American Journal of Agricultural Economics* 61: 228–237.
Glazer, A. 1981. Advertising, information and prices: A case study. *Economic Inquiry* 19: 661–671.
Kwoka, J.E. Jr. 1984. Advertising and the price and quality of optometric services. *American Economic Review* 74: 211–216.
Milyo, J., and J. Waldfogel. 1999. The effect of price advertising on prices: Evidence in the wake of 44 Liquormart. *American Economic Review* 89: 1081–1096.

Advisers

John Wood

Since olden times, princes, powers and potentates with widely differing economic backgrounds have availed themselves of advisers and advisory services of various kinds. Advisers sometimes assumed such positions of power and influence that, due either to their expertise or to their influence over the decisions of their employers, they became virtual rulers of a State. Father Joseph's influence over Cardinal Richelieu was such that he became known as the 'Eminence grise', a term that became part of English usage. At the other end of the scale, monarchs would sometimes make use of court jesters and buffoons to advise them on public opinion judged from the reactions to the jesters' gibes and jokes.

Advisory opinions are frequently sought in English Common Law practice, as well as in the medical field. Advisory services are also frequently described as counselling services. In Embassies the title of Counsellor is generally used to indicate the most senior staff member after the Ambassador. The word in fact, derives from the French 'conseiller' or adviser. This is an indication of the importance attached to such functions in established diplomatic practice.

In more recent times, advisers have been better known in the economic field. A number of governments, including that of the United Kingdom have employed Economic Advisers. This practice has been extended, particularly in the postwar years, to the international scene, where various governments of the industrialized countries have sent economic or other technical advisers to dependent territories and to newly independent States at the latters' request. The various technical or Specialized Agencies of the United Nations system are predominantly purveyors of advisory services. These advisers, more often referred to as 'experts' are drawn from a wide range of member states of the UN and recruited on the basis of their specialized competence and may remain in overseas postings for extended periods. Consultative services are also provided by these agencies in the form of seminars, workshops and technical meetings, which could be described as collective advisory services. Shorter term consultant missions, generally under six months duration, are the usual vehicle for specific problems that can be resolved by such technical advice.

The effectiveness of advisers is, naturally, very much dependent on the status that their employers

accord them and the level in a given hierarchy at which they have to work, and on their ability to make their views heard and respected. Much will also depend on the role that the adviser is implicitly expected to play. One employer may, for instance, be a genuine querrent in search of expertise in a field which may be unfamiliar to him, another may simply be using the adviser as a 'presence' by means of which he is able to lend greater credence to his own proposals or views. More frequently though, the adviser will be expected to provide outside opinions on a range of topics, not all of which may be within his specific range of competence or of his job description. The mere fact, however, of being able to express a reasonably unbiased opinion or of coming to conclusions by approaches different from those normally taken by his employer, is in itself an important contribution to the decision-making process of his employer.

Whatever role the adviser may play, it is important to stress the underlying principle behind the majority of such advisory positions, namely, that it provides a means of sharing or of diluting responsibility without any loss of authority on the part of the adviser's employer for decision-making.

Decision-makers, be they heads of state or junior managers, may be faced with a situation in which they realize that some possibly unpopular or risky decisions need to be taken, the results of which cannot be clearly predicted. In such cases an adviser would base his counsel on his own analyses of the problem. Should this agree in general terms with the employers' own views and inclinations, a decision would naturally be taken accordingly. Should the results of such advice, say for some unforeseen reason, turn out to be politically or economically disadvantageous, the employer can readily salvage his reputation by letting it be known that his decision was taken on the basis of the best available national or international advice. Should that not be sufficient to head off criticism, the adviser can be dismissed, carrying the blame for the erroneous decision. This would then enable the employer to take another course without undue damage to his own position or status. On the other hand, should the advice given to the employer be contrary to his views, he had the choice of either throwing it and the adviser out, or of allowing it to go forward into action with the responsibility for the consequences falling directly on the adviser. A successful outcome under such circumstances would then redound to the credit of both the employer and the adviser.

From this it can be seen that an important prerequisite for an adviser is an ability to use foresight to correctly forecast developments that are likely to flow from the advice given. Forecasting in a limited or specific isolated technical field is not particularly hazardous, but as soon as more complex issues relating, say, to economic policy, macroeconomic projections, or futures scenarios are considered, forecasting on the basis of often multiple variables, becomes, even with the aid of computer technology, much more prone to error. It should be borne in mind that under such circumstances the adviser represents and bases his advice on 'science', that is technical expertise, while his employer – generally a policy- or decision-maker – represents action. In practice the interface between these often becomes blurred. If the 'scientist' adviser limits his actions to factual data carrying no value judgements whatsoever, he may be failing in his assigned role of giving pertinent advice. Yet, if he draws too many conclusions from his data he may be usurping the prerogatives of his employer, a process that can lead to the adviser becoming an *éminence grise*.

The forecasting ability of an adviser is also conditioned by the level at which he is placed in a hierarchy. Complex issues of policy or of trend forecasting are nearly always contingent upon other related factors. If there are no clear guidelines from the level above that at which the adviser is working, he will not be able to provide much more than theoretical hypotheses. Again, if at a higher level he is unable to have access to other sectors of activity that may impinge on his own, his advice will be of limited value. Actively to seek out such information might be considered by the other sectors as an infringement or interference. This is particularly so in the case of governmental departments which tend to be rather rigidly

hierarchized and jealous of their prerogatives. In order to provide an adviser with the freedom to range across such boundaries, they are often assigned to planning departments or Ministries. This practice is common in the case of internationally assigned economic or policy advisers. A disadvantage arising from this expedient is that the adviser becomes further removed from the action side of his role and more involved in theory and the elaboration of more utopian proposals that may not be realized.

An adviser can often be placed in a position where his advice has been overriden for, say, political or extraneous security reasons of which he may not have been cognizant, and yet retain the respect and support of his employer. He would then most likely be asked to assess the consequences of decisions he had not worked on, or perhaps not even envisaged. This situation occurs in Civil Services in respect of Ministerial decisions and requires both flexibility and a complete detachment on the part of the adviser from the implementation of his advice. This is a quality particularly valuable to decision-makers who are sometimes obliged for quite 'unscientific' reasons to embark on actions they have earlier condemned. The adviser could be relied upon to continue to provide unbiased technical advice based on new sets of parameters and probabilities.

Advisers can play important, and sometimes determinant roles in national and international affairs but remain as a general rule anonymous. It is consequently difficult for historians to assess their true role and contribution. When they have entered the pages of history it has often been for the wrong kind of notoriety. In the case of internationally assigned advisers this anonymity is subsumed into the collective efforts of the various organizations working in this field which in the case of intergovernmental organizations function as an international civil service.

See Also

► Courcelle-Seneuil, Jean Gustave (1813–1892)
► Forecasting

Bibliography

Jöhr, W.A., and H.W. Singer. 1955. *The role of the economist as official adviser*. London: George Allen & Unwin.

Affine Term Structure Models

Michael W. Brandt and David A. Chapman

Abstract

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

Keywords

Affine term structure models; Arbitrage opportunities; Bonds; Brownian motion; Continuous-time models; Expectations hypothesis; Generalized method of moments; Itô integral; Liquidity preference; Maximum likelihood; Multifactor models; Ornstein–Uhlenbeck processes; Preferred habitat theory; Principal components; Single-factor models; State price deflators; Term premiums; Term structure of interest rates; Zero-coupon bonds

JEL Classification

D4; D10

The term structure of interest rates refers to the relationship between the yields-to-maturity of a set of bonds and their times-to-maturity. It is a simple descriptive measure of the cross-section of bond prices observed at a point in time. An affine term structure model hypothesizes that the term structure of interest rates at any point in time

is a time-invariant linear function of a small set of common state variables or factors. Once the dynamics of the state variables and their risk premiums are specified, the dynamics of the term structure are determined.

For the term structure of interest rates to be meaningful, the bonds being compared must have similar risk and payout characteristics. The literature we examine in this article focuses on the term structure of default-risk free nominal bonds that make a single payment at a pre-specified future date – so-called zero-coupon bonds. The models described below can be applied to other types of bonds, but zero-coupon bonds are particularly important because they represent the fundamental discount rates embedded in all financial claims that make payments through time.

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

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

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

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

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

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

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

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

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

As Eq. (1) clearly shows, pricing zero-coupon default-risk-free bonds boils down to specifying a

model for the dynamics of the short rate under the risk-neutral measure. In choosing models for r_t, there are two paramount considerations: (*a*) a flexible specification that does a reasonable job of capturing the dynamics of proxies for the short rate (since r_t itself is unobservable), and (*b*) a specification that yields a convenient form for the bond prices that are the ultimate objects of interest.

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

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

Single-Factor Models

In a single-factor affine model, the determinant of bond prices is the short rate itself. The model is constructed by specifying a continuous-time process for the short rate and a form of the risk premium function. As Cox et al. (1985) note, these choices must be mutually consistent in order to avoid accidentally introducing arbitrage opportunities into a (supposedly) arbitrage-free model. The fundamental building blocks of all affine models are the single-factor models due to Vasicek (1977) and Cox et al. (1985) (hereafter CIR).

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

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

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

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

where

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

and

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

Equation (5) is the first statement of an exponential-affine pricing function. It implies a simple structure where continuously compounded yields are Gaussian with constant volatility. The term structure of forward rates implied by this simple model can assume most (but not all) of the commonly observed shapes of the term

structure. In particular, the term structure of forward rates can be upward sloping, downward sloping, or humped shaped, although the model cannot generate an inverted humped shape. Since prices at all maturities are driven by a single stochastic factor, this model implies that all yield levels are perfectly correlated. In the data, yield levels are very highly, but not perfectly, correlated.

In the single-factor CIR term structure model, the short rate evolves as

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

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

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

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

where $\gamma \equiv \sqrt{(\kappa+\lambda_1)^2 + 2\sigma^2}$. The CIR model can generate the most common shapes of the term structure, but it still implies that all yield levels are perfectly correlated.

The Vasicek and CIR models are the most common forms of single-factor affine models, but Duffie and Kan (1996) provide the conditions on the drift, diffusion, and risk premium functions of a short rate specification, like (4) or (8), that ensure that the bond pricing function is exponential-affine under the risk neutral measure. In particular, a pricing function of the form of (5) will follow if

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

and

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

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

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

Multifactor Models

If single-factor models are insufficient to explain the observed term structure, then how many factors are needed and what are the dynamics of these factors? The common answer to the first question

is provided by the analysis of Litterman and Scheinkman (1991). Using a simple principal components approach, they argue that three factors, extracted from bond yields or returns themselves, can explain well over 95% of the variation in weekly changes of US Treasury bond prices, for maturities of up to 18 years. The answer to the second question – in the most general form consistent with an exponential-affine pricing function – is provided by Dai and Singleton (2000) and extended by Duffee (2002).

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

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

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

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

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

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

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

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

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

$$P_t(\tau) = \exp\left[A(\tau) + B(\tau)' Y_t\right]. \tag{16}$$

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

$$\frac{dA(\tau)}{d\tau} = -\theta K' \; B(\tau) + \frac{1}{2}\sum_{i=1}^{N}\left[\Sigma' B(\tau)\right]_i^2 \alpha_i - \delta_0 \tag{17}$$

and

$$\frac{dB(\tau)}{d\tau} = -K' \; \beta(\tau) + \frac{1}{2}\sum_{i=1}^{N}\left[\Sigma' B(\tau)\right]_i^2 \beta_i - \delta. \tag{18}$$

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

$$\Lambda_t = \sqrt{S_t}\lambda_0 + \sqrt{S_t^-}\lambda Y_t, \tag{19}$$

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

$$S_t^-(ii) = \begin{cases} \left(\alpha_i + \beta_i' Y_t\right)^{-1/2}, & \text{if } \inf\left(\alpha_i + \beta_t' Y_t\right) > 0; \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

The first term in (19) is a straightforward generalization of the single-factor risk premium

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

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

Dai and Singleton (2002), Duffee (2002) and Brandt and Chapman (2006) fit multifactor affine term structure models to more than 25 years of monthly US bond data. Each paper considers the ability of different versions of $\mathbb{A}_m$ (3) models to both explain the rejections of the expectations hypothesis and to provide accurate forecasts of future yields. Both Dai and Singleton (2002) and Brandt and Chapman (2006) find that a Gaussian version (an $\mathbb{A}_0$ (3) model) can rationalize the risk premiums dynamics revealed by expectations hypothesis tests. Duffee (2002) demonstrates that an $\mathbb{A}_0$ (3) model with the expanded risk premium specification of (19) can produce more accurate yield forecasts than a random walk benchmark model.

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

Concluding Remarks

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

See Also

- Arbitrage
- Continuous and Discrete Time Models
- Finance
- Finance (new developments)
- Linear Models
- Markov Processes
- Term Structure of Interest Rates
- Wiener Process

Bibliography

Ang, A., and M. Piazzesi. 2003. A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *Journal of Monetary Economics* 50: 745–787.

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

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

Campbell, J. 1986. A defense of traditional hypotheses about the term structure of interest rates. *Journal of Finance* 41: 183–193.

Campbell, J., and R.J. Shiller. 1991. Yield spreads and interest rate movements: A bird's eye view. *Review of Economic Studies* 58: 495–514.
Chan, K., G. Karolyi, F. Longstaff, and A. Sanders. 1992. An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance* 47: 1209–1227.
Chapman, D., and N. Pearson. 2001. Recent advances in estimating term structure models. *Financial Analysts Journal* 57: 77–95.
Cox, J.C., J. Ingersoll Jr., and S. Ross. 1981. A re-examination of traditional hypotheses about the term structure of interest rates. *Journal of Finance* 36: 769–799.
Cox, J., J. Ingersoll Jr., and S. Ross. 1985. A theory of the term structure of interest rates. *Econometrica* 53: 385–408.
Dai, Q., and K. Singleton. 2000. Specification analysis of affine term structure models. *Journal of Finance* 55: 1943–1978.
Dai, Q., and K. Singleton. 2002. Expectation puzzles, time-varying risk premia, and affine models of the term structure. *Journal of Financial Economics* 63: 415–441.
Dai, Q., and K. Singleton. 2003. Term structure dynamics in theory and reality. *Review of Financial Studies* 16: 631–678.
Duffee, G. 2002. Term premia and interest rate forecasts in affine models. *Journal of Finance* 57: 405–443.
Duffie, D. 2001. *Dynamic asset pricing theory*, 3rd ed. Princeton: Princeton University Press.
Duffie, D., and R. Kan. 1996. A yield factor model of interest rates. *Mathematical Finance* 6: 379–406.
Durham, G. 2003. Likelihood-based specification analysis of continuous-time models of the short-term interest rate. *Journal of Financial Economics* 70: 463–487.
Fama, E., and R. Bliss. 1987. The information in long maturity forward rates. *American Economic Review* 77: 680–692.
Litterman, R., and J. Scheinkman. 1991. Common factors affecting bond returns. *Journal of Fixed Income* 3: 54–61.
Piazzesi, M. 2005. Affine term structure models. In *Handbook of financial econometrics*, ed. Y. Aït-Sahalia and L. Hansen. Amsterdam: North-Holland.
Vasicek, O. 1977. An equilibrium characterization of the term structure. *Journal of Financial Economics* 5: 177–188.

Affirmative Action

Harry J. Holzer and David Neumark

Abstract

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

Keywords

Affirmative action; Black–white wage differences; Efficiency; Labour market discrimination; Labour-market institutions; National Educational Longitudinal Study (NELS); Redistribution; Women's work and wages

JEL Classifications

J15; J16; J7

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

Legal Underpinnings and Controversies

The roots of affirmative action in employment lie in a set of Executive Orders issued by US Presidents since the 1960s. Executive Order 10925 (issued in 1961) introduced the term 'affirmative action', encouraging employers to take action to ensure non-discrimination. Executive Order 11246 (1965) required federal contractors and subcontractors (currently, with contracts of $50,000 or more) to identify underutilized minorities, to assess availability of minorities, and if available, to set goals and timetables for reducing the underutilization. Executive Order 11375 (1967) extended this requirement to women.

Federal contractors may be sued and barred from contracts if they are judged to be discriminating or not pursuing affirmative action, although this outcome is rare (Stephanopoulos and Edley 1995). But affirmative action is not just limited to contractors; it can be imposed on non-contractor employers by courts as a remedy for past discrimination, and it can be undertaken voluntarily by employers.

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

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

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

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

Affirmative action remains vastly more controversial than anti-discrimination activity, even though the distinctions between them are clearer in theory than in practice (Holzer and Neumark 2000a). The critics of affirmative action argue that it transfers jobs, university admissions, and business contracts to minorities and women at the expense of white males who might be more qualified and therefore more deserving. If so, it might constitute a form of 'reverse discrimination' against white males, which could be both inefficient and unfair. In contrast, the supporters of affirmative action claim that extra efforts beyond just the removal of explicit discrimination are necessary to overcome the many inherent disadvantages that minorities and women face in universities, the labour market, and the business sector. On this view, affirmative action is necessary for equal opportunity (or 'fairness'), and would not necessarily reduce efficiency. Indeed,

it might even raise overall efficiency by making available a wider pool of talent on which businesses and universities could draw, or because diversity itself has positive impacts.

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

Redistributive Effects

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

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

Furthermore, analyses of micro-level data on applications and admissions by Kane (1998) and by Long (2004) suggest somewhat modest effects of affirmative action on overall admissions of minorities, but both studies suggest that the magnitudes rise with the overall level of scores at universities. Using data from the High School and Beyond Survey, Kane found significant racial differences in admissions (conditional on test scores and many other personal characteristics) only in the top quintile of colleges and universities by test scores. Long, using data from the National Educational Longitudinal Study (NELS), found significant effects on admissions in all quintiles. But the magnitudes of these differences were not large in absolute terms – the probability that minorities are accepted at their top choice would decline by less than 2 percentage points (14.7 per cent against 16.4 per cent) in the aggregate and about 2.5 percentage points in the top quintile in the absence of affirmative action.

That affirmative action is more important as college quality rises is further established by Bowen and Bok (1998), who find quite large effects at a set of the most prestigious colleges and universities. Indeed, their work suggests that

admissions rates among minorities at these schools would fall from 42 per cent to 13 per cent if affirmative action were abolished, a view consistent with the initial effects of Proposition 209 in California on admissions at Berkeley. The magnitudes of racial preferences in admissions in a variety of graduate programmes are also fairly large (Attiyeh and Attiyeh 1997; Davidson and Lewis 1997), while gender preferences are much more modest.

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

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

Efficiency and Performance Effects

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

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

What does the empirical evidence on the efficiency and performance of affirmative action beneficiaries show? One approach is to look at measures of individual employee credentials or performance, by race and/or sex, to see whether affirmative action generates major gaps in performance between white males and other groups. An earlier paper (Holzer and Neumark 1999) compares a variety of measures of employee credentials and performance, where the former include educational attainment (absolute levels and those relative to job requirements), and the latter include wage or promotion outcomes as well as a subjective performance measure across these groups. The study inquired whether observed gaps in

credentials and performance between white males and females or minorities are larger among establishments that practice affirmative action in hiring than among those that do not. The results indicated virtually no evidence of weaker credentials or performance among females in the affirmative action sector, relative to those of males within the same racial groups. In comparisons between minorities and whites, there was clear evidence of weaker educational credentials among the former group, but relatively little evidence of weaker performance.

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

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

A variety of other studies have been undertaken within specific sectors of the workforce, where it is easier to define employee performance. Among the sectors that have been studied are police forces (Carter and Sapp 1991), physicians (Davidson and Lewis 1997), and university faculties (Kolpin and Singell 1996). The results of these studies again show no evidence of weaker performance among women, and generally limited evidence of weaker performance among minorities. In contrast, there is evidence of potential social benefits from affirmative action in the medical sector, as minority doctors appear more likely to locate in poor neighbourhoods and treat minority or low-income patients.

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

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

The more challenging question is whether affirmative action actually hurts minority students by admitting them to colleges and universities for which some of them are unqualified, generating a poor 'fit' between them and the colleges or universities that they attend that may actually lead to worse outcomes. Sander (2004) claims to show evidence that affirmative action in law schools worsens outcomes for blacks, although this conclusion is disputable. Conversely, dropout rates of minorities at the most prestigious institutions are

generally lower than elsewhere (Bowen and Bok 1998). More decisive evidence on this question requires adequate comparison with counterfactuals of what would be observed absent affirmative action.

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

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

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

See Also

- ▶ Black–White Labour Market Inequality in the United States
- ▶ Labour Market Institutions

A

Bibliography

Antonio, A., et al. 2004. Effects of racial diversity on complex thinking in college students. *Psychological Science* 15: 507–510.

Athey, S., C. Avery, and P. Zemsky. 2000. Mentoring and diversity. *American Economic Review* 90: 765–786.

Attiyeh, G., and R. Attiyeh. 1997. Testing for bias in graduate school admissions. *Journal of Human Resources* 32: 524–548.

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

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

Bowen, W., and D. Bok. 1998. *The shape of the river*. Princeton: Princeton University Press.

Carter, D., and A. Sapp. 1991. *Police education and minority recruitment: The impact of a college requirement*. Washington, DC: Police Executives Research Forum.

Coate, S., and G. Loury. 1993. Will affirmative action policies eliminate negative stereotypes? *American Economic Review* 83: 1220–1240.

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

Davidson, R., and E. Lewis. 1997. Affirmative action and other special considerations admissions at the University of California, Davis School of Medicine. *Journal of the American Medical Association* 278: 1153–1158.

Dickens, W., and T. Kane. 1999. Racial test score differences as evidence of reverse discrimination: Less than meets the eye. *Industrial Relations* 28: 331–363.

Holzer, H., and D. Neumark. 1999. Are affirmative action hires less qualified? *Journal of Labor Economics* 17: 534–569.

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

Holzer, H., and D. Neumark. 2000b. What does affirmative action do? *Industrial and Labor Relations Review* 53: 240–271.

Kane, T. 1998. Racial preferences and higher education. In *The black-white test score gap*, ed. C. Jencks and M. Phillips. Washington, DC: Brookings Institution.

Kolpin, V., and L. Singell. 1996. The gender composition and scholarly performance of economics departments: A test for employment discrimination. *Industrial and Labor Relations Review* 49: 408–423.

Leonard, J. 1990. The impact of affirmative action regulation and equal opportunity law on employment. *Journal of Economic Perspectives* 4: 47–63.
Long, M. 2004. Race and college admissions: An alternative to affirmative action? *Review of Economics and Statistics* 86: 1020–1033.
Neumark, D., and R. Gardecki. 1998. Women helping women? Role model and mentoring effects on female Ph.D. students in economics. *Journal of Human Resources* 33: 220–246.
Sander, R. 2004. A systematic analysis of affirmative action in American law schools. *Stanford Law Review* 57: 368–483.
Stephanopoulos, G., and C. Edley. 1995. Review of federal affirmative action programs. Unpublished White House document. http://clinton4.nara.gov/textonly/WH/EOP/OP/html/aa/aa-lett.html. Accessed 25 Aug 2005.

Aftalion, Albert (1874-1956)

Joseph Halevi

Aftalion was a Bulgarian-born French economist. He taught at the University of Lille and later at the University of Paris (Villey 1968). His works include a study on Sismondi (1899), a treatise on crises of overproduction (1913), a critique of socialism (1923), two books on monetary issues (1927a, 1948) and several writings on issues related to international trade and the balance of payments (1937). His international reputation is mostly due to the 1913 work on overproduction, a summary of which exists in English (1927b).

Aftalion's approach to the problem of the trade cycle and overproduction is centred on the time lag between an expected increase in the demand for consumption goods and the production of equipment needed to generate the additional consumption goods. For this reason Aftalion has been considered as being among the inventors of the Accelerator Principle. However, his analysis differs significantly from the contemporary theories of the trade cycle based on such a principle. In those theories the Accelerator explains fluctuations in the investment component of effective demand without establishing any connection with the behaviour of prices. For Aftalion, by contrast, the time required to obtain the extra amount of equipment necessary to produce the additional consumption goods is a basic ingredient to portray the link between fluctuation in output and changes in prices. His argument, based on purely intuitive grounds, runs as follows.

An expected expansion in consumption demand will lead to larger orders by wholesale traders. Since no unused capacity is assumed more machinery will be needed, the demand for which will be propogated to all stages of production. Capitalists are assumed to plan their output on the basis of current prices. Yet, the additional demand of capital goods and raw materials cannot be immediately satisfied. Hence prices will rise in these two sectors, and eventually in the consumption goods sector as well. When the new investment projects are finished and equipment is delivered, prices begin to fall. Entrepreneurs will cut current orders but deliveries due to past investment decisions will continue, thereby reducing price and orders further. This distinction between orders and deliveries influenced Kalecki's approach to the theory of economic fluctuations.

Aftalion did not produce a theory of output because he did not attempt any explanation of the adjustment of capacity to demand. Furthermore, it is not clear whether prices of consumption goods increase because of increases in the price of raw materials or because of the expansion of demand. Indeed, since no spare capacity exists, consumption goods prices should be sensitive to changes in demand. It follows, therefore, that Aftalion's assumption of a time lag between changes in raw material prices and those of consumption goods is theoretically confusing.

See Also

► Acceleration Principle

Selected Works

1899. *L'oeuvre économique de Sismonde de Sismondi*. Paris: Pedone.
1913. *Les crises périodiques de surproduction*, 2 vols. Paris: Rivière.

1923. *Les fondements du socialisme: étude critique*. Paris: Rivière.
1927a. *La valeur de la monnaie dans l'économie contemporaine*, vol. I. Paris: Sirey.
1927b. The theory of economic cycles based on the capitalistic technique of production. *Review of Economic Statistics* 9:165–170.
1937. *L'équilibre dans les relations économiques internationales*. Paris: Domat-Montchrestien.
1948. *La valeur de la monnaie dans l'économie contemporaine,* vol. 2. Paris: Sirey.

Bibliography

Villey, D. 1968. Aftalion, A. In International encyclopedia of the social sciences. New York: Macmillan.

Ageing Populations

Robert L. Clark

Population ageing is represented by an increase in the relative number of older persons in a population and is associated with an increase in the median age of the population. The age structure of a population is determined by its mortality, fertility, and net migration experience. Although life tables and survivorship rates date from the 17th century, the development of mathematical demography is essentially a 20th-century innovation. The techniques of mathematical demography can be used to show how the age structure of a population changes with alternative transition rates.

The importance of these transition rates is shown by the observation that in the absence of migration two arbitrarily chosen populations that are subjected to identical fertility and mortality rates will ultimately generate the same age structure. Thus, as Coale (1972, p. 3) noted, populations gradually 'forget' the past in as far as their age compositions are concerned. Of course, the population age structure may echo past irregularities for several generations before these echo effects disappear (Easterlin 1980).

Population projections illustrate that declining fertility produces population ageing, so do decreases in mortality rates; however, fertility changes dominate the age structure of a population. For example, even if man were to become immortal, high fertility rates would produce a relatively young population. Migration can modify the age composition of a population, but non-sustained migration will have only a transitory effect on the age distribution of a population unless the migration also alters the prevailing patterns of fertility and mortality (Keyfitz 1968, p. 94).

Concern for the economic implications of ageing populations is essentially a 20th-century phenomenon. Populations with low life expectancies and high fertility rates will have only small fractions age 65 and older. For most of human history, these were typical population characteristics. Therefore, little attention was devoted to the macroeconomic implications of ageing. In summarizing economic thinking prior to the 20th century, Hutchinson (1967, p. 346) concludes that because the typical population age structure contained relatively few persons over age 65, not much attention was given to the ratio of workers to the total population. In most economic analysis, the population was simply assumed to be equivalent to labour supply.

Declining population growth occurred in Western Europe in the early part of the 20th century. The resulting ageing of populations began to attract attention. Economists focused their analyses on age structure ratios, such as the number of dependent persons (the young and the old) divided by the number of persons in the population or by the number of persons of working age. Much of the research examining the economic implications of ageing populations assesses the effects of changes in these dependency ratios or similar population ratios.

Dependency ratios are used to measure the relative productive potential of a population. The old-age dependency ratio generally measures the number of elderly persons at or above a certain age, say 65, divided by the number of persons of working age, say 16–64. This ratio has been widely used in economic analysis to measure the number of retired dependent persons per active member of the labour force. The old-age

dependency ratio is used to illustrate the transfer of output from workers that is necessary to support retirees. This ratio rises with population ageing.

There are several problems concerning the economic interpretation of the old-age dependency ratio. First, if population ageing follows from reduced fertility, the total dependency ratio (youths plus elderly) may fall even as the old-age ratio is rising. The total cost to society of supporting the dependent populations will depend on the relative costs of maintaining the two dependent populations and the transfer mechanisms that are developed within the economic system (Sauvy 1969, pp. 303–19). Second, the age-based dependency ratios are not perfect proxies for the ratio of inactive to active persons. Recently, some analysts have attempted to incorporate labour force participation into the dependency-ratio framework. Of course, over time, participation rates and the meaning of dependency may change. Third, significant compositional changes may occur within the elderly, youth, and working age populations. These changes have economic effects that may be as important as effects of changes in the dependency ratio itself (Clark and Spengler 1980).

The cost of national pension systems rises with population ageing because a greater fraction of the population is receiving benefits and a smaller fraction is working and paying taxes to support the system (Munnell 1977). This relationship has become one of the principal public policy issues associated with population ageing. The funding of pensions and the economic impact of alternative funding methods also has been subject to considerable examination. Feldstein (1974) argued that the pay-as-you-go financing of the US Social Security System substantially reduced the national savings rate. Subsequent research has produced a series of conflicting findings on this issue.

The growth of national pension systems has drawn attention to retirement ages. The impact of population ageing on pension funding requirements is exacerbated if the age of withdrawal from the labour force declines. During the past century, labour force participation rates of the elderly have fallen and the interaction of earlier retirement and population ageing has produced significant increases in income transfers to the elderly.

The changing age structure of a population may also alter the equilibrium unemployment rate and the average level of productivity in a society. Layoff and quit rates are a decreasing function of age. Since employment stability increases with age, national unemployment rates tend to decline with population ageing. Some attention has been given to the effect of ageing on productivity with emphasis on the ageing of the labour force and the ensuing slower rate of introduction of new human capital into the production process. The ability of older workers to maintain production standards has also been questioned. Data limitations preclude a definitive answer to the shape of the age-productivity profile. The macroeconomic significance of population ageing on national productivity depends on individual age-specific productivity, and any ensuing changes in investment, consumption, and savings behaviour. The net effect of these factors is unclear.

The effect of population ageing on national savings and therefore the rate of economic growth depends on age-specific savings rates and the age structure changes that occur as the population ages (Kelley 1973). Although ageing of individuals tends to reduce their savings in old age, population ageing typically is associated with an increase in the fraction of the population in the high savings years and thus tends to stimulate increased saving and investment. The net effect of ageing on savings and growth will also depend on the cause of the population ageing. If population ageing results from slowing population growth, then the economic response to population size and rate of population growth will be observed simultaneously with the ageing effect. In general, the independent effect of population ageing will not be a major factor influencing future economic growth and development.

See Also

- ▶ Declining Population
- ▶ Demographic Transition
- ▶ Social Security

► Stable Population Theory
► Stagnation

Bibliography

Clark, R., and J. Spengler. 1980. *Economics of individual and population ageing*. Cambridge: Cambridge University Press.
Coale, A. 1972. *The growth and structure of human populations: A mathematical investigation*. Princeton: Princeton University Press.
Easterlin, R. 1980. *Birth and fortune*. New York: Basic Books.
Feldstein, M. 1974. Social security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy* 82(5): 905–926.
Hutchinson, E.P. 1967. *The population debate*. Boston: Houghton-Mifflin.
Kelley, A. 1973. Population growth, the dependency rate and the pace of economic development. *Population Studies* 27(3): 405–414.
Keyfitz, N. 1968. *Introduction to the mathematics of population*. Reading: Addison-Wesley.
Munnell, A. 1977. *The future of social security*. Washington, DC: The Brookings Institution.
Sauvy, A. 1969. *General theory of population*. New York: Basic Books.

Agency Costs

Clifford W. Smith Jr.

In the traditional analysis of the firm, profit maximization is assumed, subject to the constraints of a technological production function for transforming inputs into output. Optimum production solutions are characterized in terms of the equality between the ratio of marginal products of inputs and the ratio of input prices. While this analysis has provided valuable insights in understanding certain aspects of choices by firms, it completely ignores others having to do with the process through which the inputs are organized and coordinated. In essence, the traditional economic analysis treats the firm as a black box in this transformation of inputs into output. Rarely are questions raised such as: Why are some firms organized as individual proprietorships, some as partnerships, some as corporations, and others as cooperatives or mutuals? Why are some firms financed primarily by equity and others with debt? Why are some inputs owned and others leased? Why do some industries make extensive use of franchising while others do not? Why do some bonds contain call provisions, convertibility provisions, or sinking fund provisions while others do not? Why are some executives compensated with salary while others have extensive stock option or bonus plans? Why do some industries pay workers on a piece-rate basis while others pay at an hourly rate? Why do some firms employ one accounting procedure while others choose alternate procedures? To answer such questions requires the economic analysis of contractual relationships. Agency Theory provides a framework for such an analysis.

An agency relationship is defined through an explicit or implicit contract in which one or more persons (the principal(s)) engage another person (the agent) to take actions on behalf of the principal(s). The contract involves the delegation of some decision-making authority to the agent. Agency costs are the total costs of structuring, administering, and enforcing such contracts. Agency costs, therefore, encompass all contracting costs frequently referred to as transactions costs, moral hazard costs, and information costs.

Jensen and Meckling (1976) break down agency costs into three components: (1) monitoring expenditures by principal, (2) bonding expenditures by the agent, and (3) the residual loss. Monitoring expenditures are paid by the principal to regulate the agent's conduct. Bonding expenditures are made by the agent to help assure that the agent will not take actions which damage the principal or will indemnify the principal if the prescribed actions are undertaken. Hence, monitoring and bonding costs are the out-of-pocket costs of structuring, administering, and enforcing contracts. The residual loss is the value of the loss by the principal from decisions by the agent which deviate from the decisions which would have been made by the principal if he had the same information and talents as the agent. Since it is profitable to invest in policing contracts only to the point where the reduction in the loss from

non-compliance equals the incremental costs of enforcement, the residual loss is the opportunity loss when contracts are optimally, but incompletely enforced.

Jensen and Meckling (1976) point out that agency problems emanating from conflicts of interests are common to most cooperative endeavours whether or not they occur in the hierarchial manner implied in the principal–agent analogy. But, with the elimination of the difference between principal and agent, the distinction between monitoring and bonding costs is also lost; so, total agency costs are out-of-pocket costs plus the opportunity cost or residual loss.

It is crucial to recognize that the contracting parties bear the agency costs associated with their interaction and therefore have incentives to structure contracts to reduce agency costs wherever possible. Within the contracting process, incentives exist for individuals to negotiate contracts specifying monitoring and bonding activities so long as their marginal cost is less than the marginal gain from reducing the residual loss. Specifically, the contracting parties gain from forecasting accurately the actions to be undertaken and structuring the contracts to facilitate the expected actions. For example, with competitive and informationally efficient financial markets, unbiased estimates of agency costs should be included in the prices of securities when they are initially offered (as well as at any future date). This mechanism provides incentives to structure contracts and institutions to lower agency costs. Hence, in the absence of the usual externalities, the private contracting process produces an efficient allocation of resources.

Jensen (1983) describes two approaches to the development of a theory of agency which he labels the 'positive theory of agency' and the 'principal-agent' literatures. Both approaches examine contracting among self-interested individuals and both postulate that agency costs are minimized through the contracting process; thus, both address the design of Pareto-efficient contracts. However the approaches diverge at several junctures. The principal-agent literature generally has a mathematical and non-empirical orientation and concentrates on the effects of preferences and asymmetric information (for example, Harris and Raviv 1978; Holmstrom 1979; Ross 1973; Spence and Zeckhauser 1971). The positive agency literature generally has a non-mathematical and empirical focus and concentrates on the effects of the contracting technology and specific human or physical capital (for example, Fama and Jensen 1983a, b; Jensen and Meckling 1976; Myers 1977; Smith and Warner 1979).

The investigation of agency costs has provided a deeper understanding of many dimensions of complex contractual arrangements, especially the modern corporate form. One can better understand the variation in contractual forms across organizations by studying the nature of the agency costs in alternative contractual arrangements. For example, Fama and Jensen (1983a) examine the nature of residual claims and the agency costs of separation of management and riskbearing to provide a theory of the determinants of alternative organizational forms. They argue that corporations, proprietorships, partnerships, mutuals and non-profits differ in the manner they trade off the benefits of risk-sharing with agency costs.

Agency cost analysis has been employed to examine the choice of organizational structure in the insurance (Mayers and Smith 1981, 1986) and thrift industries (Smith 1982; and Masulis 1986). It has also been employed to examine the determinants of the firm's capital structure (Jensen and Meckling 1976; Myers 1977); the provisions in corporate bond contracts (Smith and Warner 1979); the determinants of corporate leasing policy (Smith and Wakeman 1985) and franchise policy (Brickley and Dark 1987); the incentives for the development of a hierarchical structure within organizations (Zimmerman 1979; Fama and Jensen 1983b); and the determinants of corporate compensation policy (Smith and Watts 1982). Finally, the analysis of agency costs has played a central role in the development of a positive theory of the choice of accounting techniques (Watts and Zimmerman 1986).

Agency analysis has also afforded a different perspective in assessing the implications of observed contractual provisions. For example, typical discussions of mortgage loan provisions suggest that escrow accounts and limitations on

renting the property are included in the loan contract for the benefit of the lender. However, if there is competition among lenders, these benefits must be reflected in compensating differentials in other loan terms, such as lower promised interest rates. If in addition, the rates on other securities are not affected by changes in the terms of this contract, then all of the benefits of these convenants must ultimately accrue to the borrower, not the lender.

See Also

- ▶ Incentive Contracts
- ▶ Principal and Agent (i)
- ▶ Transaction Costs

Bibliography

Brickley, J.A., and F.H. Dark. 1987. The choice of organizational form: The case of franchising. *Journal of Financial Economics* 18: 401.

Fama, E., and M. Jensen. 1983a. Agency problems and residual claims. *Journal of Law and Economics* 26: 327–349.

Fama, E., and M. Jensen. 1983b. Separation of ownership and control. *Journal of Law and Economics* 26: 301–325.

Harris, M., and A. Raviv. 1978. Some results on incentive contracts with applications to education and employment, health insurance and law enforcement. *American Economic Review* 68: 20–30.

Holmstrom, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10(1): 74–91.

Jensen, M. 1983. Organization theory and methodology. *Accounting Review* 58: 319–339.

Jensen, M., and W. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.

Masulis, R. 1986. Changes in ownership structure: Conversions of mutual savings and loans to stock charter. *Journal of Financial Economics*.

Mayers, D., and C. Smith. 1981. Contractual provisions, organizational structure, and conflict control in insurance markets. *Journal of Business* 54: 407–434.

Mayers, D., and C. Smith. 1986. Ownership structure and control: The mutualization of stock life insurance companies. *Journal of Financial Economics* 16: 73.

Myers, S. 1977. Determinants of corporate borrowing. *Journal of Financial Economics* 5: 147–175.

Ross, S. 1973. The economic theory of agency: The principal's problem. *American Economic Review* 63: 134–139.

Smith, C. 1982. Pricing mortgage originations. *AREUEA Journal* 10: 313–330.

Smith, C., and L. Wakeman. 1985. Determinants of corporate leasing policy. *Journal of Finance* 40: 895–908.

Smith, C., and J. Warner. 1979. On financial contracting: An analysis of bond covenants. *Journal of Financial Economics* 7: 117–161.

Smith, C., and R. Watts. 1982. Incentive and tax effects of U.S. executive compensation plans. *Australian Journal of Management* 7: 139–157.

Spence, M., and R. Zeckhauser. 1971. Insurance, information, and individual action. *American Economic Review* 61: 119–132.

Watts, R., and J. Zimmerman. 1986. *Positive accounting theory*. Englewood Cliffs: Prentice-Hall.

Zimmerman, J. 1979. The costs and benefits of cost allocations. *The Accounting Review* 54: 504–521.

Agency Problems

Luca Anderlini and Leonardo Felli

Abstract

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

Keywords

Agency problems; Commitment; Common values; Continuous-time models; Contract theory; Discrete-time models; First-order approach; Incentive design; Insurance–incentives trade-off; Intertemporal incentives; Limited liability; Linear incentive schemes; Menu contracts; Noisy tasks; Non-profit organizations; Pooling equilibria; Principal and agent; Separating equilibria; Signalling; Soft incentives

JEL Classifications

D23

Within modern economic analysis, early recognition of the importance of agency problems goes back to at least Marschak (1955), Arrow (1963) and Pauly (1968). These early works are followed by the classical contributions of Mirrlees (1975/1999), Holmström (1979), Shavell (1979) and Grossman and Hart (1983).

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

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

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

When both P and $\mathcal{A}$ are risk-neutral, an agency problem also arises if the *LC binds* (and typically the *IR* does not). (If the reverse is true, then giving incentives to $\mathcal{A}$ has no cost since he does not mind risk and the *IR* binds on his *expected* payoff. In fact in this case, the 'social optimum' coincides with the 'constrained social optimum' in which a social planner can choose e, but only subject to giving the appropriate incentives to $\mathcal{A}$.) In this case in order to give $\mathcal{A}$ incentives P can pay him more when output indicates that effort is higher. This drives a wedge between P's marginal cost for increased e and its social marginal cost. This in turn dictates that he equilibrium contract will differ from the first-best, and a 'second-best' 'constrained-inefficient' outcome obtains.

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

A Simple Benchmark

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

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

world. This generates the two *LC*s $w_1 \geq 0$ and $w_0 \geq 0$.

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

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

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

Multi-tasking

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

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

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

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

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

When $p \leq 3/5$, σ_N is not informative enough. In this case $e_S = e_N = w_{S1} = w_{N1} = 0$. This has been interpreted as *no contract* being signed. The no-contract outcome obtains even though an informative contractible signal for *both* tasks is available.

Informed Principal

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

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

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

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

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

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

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

Intertemporal Incentives

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

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

$\mathcal{A}$ is paid at the end of the two periods, as a function of observed output in the two periods. The wage paid if output is $i \in \{0, 1\}$ in period F and $j \in \{0, 1\}$ in period S is denoted w_{ij}.

Neither P nor $\mathcal{A}$ discounts the future. While P is risk-neutral, $\mathcal{A}$ is risk-averse with an *exponential* utility with a constant absolute risk-aversion coefficient equal to 1/2. His effort in the two periods is perfectly substitutable. Given a wage scheme w_{ij} and effort levels e_F and e_{iS} his expected utility is

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

while P's expected payoff is

$$\begin{aligned}\Pi^{P} = &\sqrt{e_F}[\sqrt{e_{1S}}(2-w_{11}) + (1-\sqrt{e_{1S}})(1-w_{10})]\\ &+(1-\sqrt{e_F})[\sqrt{e_{0S}}(1-w_{01}) + (1-\sqrt{e_{0s}})(-w_{00})]\end{aligned}$$

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

The *IR* constraint is binding for $\mathcal{A}$ while it is not binding for P. The *IC* constraint can be subsumed in the *first order* conditions obtained by differentiating $\Pi^{\mathcal{A}}$ with respect to e_F and e_{iS} and setting these equal to 0 which are sufficient for a maximum. This way to proceed is known in the literature as taking the *first-order approach*. In the more general case considered for instance by Holmström and Milgrom (1987) this is not viable. In the simple case considered here, the first-order approach works because we are assuming that the exponent of effort variables – 1/2 in this case – plus $\mathcal{A}$'s constant absolute risk-aversion coefficient – also 1/2 in this case – sum to 1. Even in single-period agency models, whether the first-order approach is valid or not is an intricate question first uncovered by Mirrlees (1975/1999). Subsequent contributions on this topic can be found in Grossman and Hart (1983), Rogerson (1985) and Jewitt (1988). To characterize the optimal incentive scheme for the two-period problem it is useful to first consider the second period (S) sub-problem after output $i \in \{0, 1\}$ has been realized in the first period (F). These problems are obtained considering (continuation) payoffs for $\mathcal{A}$ and P given by the relevant square bracket term of $\Pi^{\mathcal{A}}$ and Π^P above, and with an *IR* constraint for $\mathcal{A}$ given by his utility level (contingent on output in F) in the solution to the two-period problem, after factoring out the common term $\{e_F/2\}$.

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

The period S sub-problems can then be plugged into the two-period problem. Viewed from period F we can think of P as offering $\mathcal{A}$ two *certainty equivalent* wages c_i for each period F output. Notice that we can write $c_i = \tilde{w}_i - \pi_i$ where $\tilde{w}_i$ is

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

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

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

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

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

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

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

Recent Developments

Since its inception the literature on agency problems and applications has grown dramatically, influencing many areas of economics ranging from development to finance. Agency theory has found a prominent place in many graduate and undergraduate programs in economics. Recent texts that provide a comprehensive treatment of the field include Salanié (2000), Laffont and Martimort (2002) and Bolton and Dewatripont (2005). Recent developments in the actual analytical framework relax some of the basic assumptions of the canonical model.

Eliaz and Spiegler (2006) and O'Donoghue and Rabin (2005) focus on the underlying *behavioural* assumptions. The first paper tackles an environment in which agents may differ in their *cognitive abilities*, which generates dynamically inconsistent behaviour. The second paper is concerned with the effect of *present bias* in the agent's preferences on the optimal incentive scheme. In both cases the optimal incentive scheme becomes more realistically 'sensitive to detail' than in the standard case.

Besley and Ghatak (2005) focus on the case of *motivated agents* in the provision of a public

good. Motivated agents do not always regard effort as a cost. This has important effects on incentive design, which in turn sheds light on the nature of non-profit organizations.

See Also

- ▶ Contract Theory
- ▶ Incentive Compatibility
- ▶ Incomplete Contracts
- ▶ Mechanism Design
- ▶ Moral Hazard

Bibliography

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

Arrow, K. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53: 941–973.

Besley, T., and M. Ghatak. 2005. Competition and incentives with motivated agents. *American Economic Review* 95: 616–636.

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

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

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

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

Hellwig, M., and K. Schmidt. 2002. Discrete-time approximations of the Holmström–Milgrom Brownian-motion model of intertemporal incentive provision. *Econometrica* 70: 2225–2264.

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

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

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

Holmström, B., and P. Milgrom. 1994. The firm as an incentive system. *American Economic Review* 84: 972–991.

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

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

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

Marschak, J. 1955. Elements for a theory of teams. *Management Science* 1: 127–137.

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

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

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

Myerson, R. 1983. Mechanism design by an informed principal. *Econometrica* 51: 1767–1798.

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

Pauly, M. 1968. The economics of moral hazard. *American Economic Review* 58: 531–537.

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

Rogerson, W. 1985. The first-order approach to principal–agent problems. *Econometrica* 53: 1357–1367.

Salanié, B. 2000. *The economics of contracts*. Cambridge, MA: MIT Press.

Shavell, S. 1979. On moral hazard and insurance. *Quarterly Journal of Economics* 93: 541–562.

Agent-Based Models

Scott E. Page

Abstract

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

ground between stark, dry rigorous mathematics and loose, possibly inconsistent, descriptive accounts.

Keywords

Agent-based models; Behavioural game theory; Central limit theorem; Complexity; Conway's game of life; Economic complexity; Emergence; Equilibrium; Interaction structures; Learning and information aggregation in networks; Mathematics and economics; Prisoner's dilemma; Rule-based behaviour

JEL Classifications

C6; D5

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

Agent-based models of economies, like real economies, consist of computational objects that interact according to rules. Agent-based modelling allows us to consider richer environments with greater fidelity than do existing techniques (Tesfatsion 1997). This increased fidelity results from the inductive nature of the modelling enterprise. When constructing an agent-based model, we are constrained only by our imagination and interest. In contrast, when constructing a mathematical model, we must always be concerned with analytic tractability. This constrains our endeavours. The set of models that one believes to be tractable is small when compared with the set of models worth exploring. Thus, the flexibility and potential for realism enlarge the set of questions economists can explore (Anderson et al. 1988; Arthur et al. 1997).

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

The Benefits of Agent-Based Models

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

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

most of the time, will come close to efficiently allocating goods and services.

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

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

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

The output flexibility of agent-based models leads some to jump to the inaccurate and unfortunate conclusion that agent-based models preclude equilibrium analysis. True, agent-based models naturally allow for dynamics, but this does not mean that they cannot attain equilibria. These equilibria are not assumed by *generated* (Epstein 2003). The generative claim that 'if you didn't grow it, you didn't show it' should be ignored at our peril. Proving that an equilibrium exists and showing that it can be attained and maintained are separate findings. But not all agent-based models generate the equilibria predicted by mathematics. They fail because attaining equilibrium often requires slow learning rates and lots of agents. Sometimes, though, they fail because the mathematics contains errors (Page and Tassier 2004).

Attaining equilibria to complement mathematical analyses (Judd 1997) is not the reason to use agent-based models. They are better suited to exploring those parts of the economy that are complex or on the boundary between complexity and equilibrium. Even critics of agent-based modelling admit the appeal of exploring complexity, but they question what we learn from individual models. Mathematical theorems *prove* results for entire classes of functions. Arrow, Debreu and McKenzie proved theorems for any convex preferences, not just for preferences derived from Cobb–Douglas utility functions. Agent-based models, at least for now, assume particular functional forms. Mathematics therefore gives us the kind of general results on which a science has traditionally been built. Agent-based models do not. This is only partly true. These critics are less than honest about the current state of our knowledge (Leombruni and Richiardi 2005). Although mathematical theorems are general and agent-based models are particular, that is not the whole story. In economics, general results are few and far between. Many papers (*a*) assume specific functional forms rendering them examples not general truths, or (*b*) consider restrictive classes of functional forms such as quasi-linear preferences, or (*c*) rely on dubious assumptions such as the monotone likelihood ratio property or independent signals.

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

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

Economic Actors as Objects

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

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

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

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

Finally, the idea of agents as objects demands explicit consideration of agent activation. In what order do the agents get called to take their action? Do they get called simultaneously or

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

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

Emergence

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

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

In economies and societies, many things emerge: prices, cities, trade patterns, information networks, and cultural norms, to name just a few (Tesfatsion and Judd 2006). These features of our world matter for economies. Cities matter. Trade networks matter. Culture matters. Social science needs ways of understanding how these things come to be as well as how they influence the performance of economic and political systems. Agent-based models offer a route to those understandings that complements our mathematical approaches.

Complexity

Agent-based models can generate complexity and allow us to explore its causes, thereby interweaving the methodology of agent-based models with the theoretical idea of complexity. The four main features of agent-based models are diverse agents, situated in an interaction structure, whose actions create interactive effects (externalities), which

adapt, evolve or learn each contribute to the level of complexity a model produces (Axelrod and Cohen 2000). These features can be thought of as choice variables. We can imagine a knob for each feature – a diversity knob, a connectedness knob, an externality knob, and a learning rate knob. The agents can be nearly homogeneous or very diverse. The space can be sparsely connected or highly connected. The interactions can be few and small or numerous and large, and the agents can adapt not at all or instantaneously. By turning these knobs, we can create complexity.

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

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

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

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

The Future of Modelling

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

Resonance provides one strong reason. Agent-based models contain people and firms embedded in interaction structures. These people and firms have conceptualizations of problems and

situations. At times, they adhere to routines. At times, they experiment.

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

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

See Also

- ▶ Behavioural Game Theory
- ▶ Learning and Information Aggregation in Networks
- ▶ Mathematics and Economics

Bibliography

Anderson, P., K. Arrow, and D. Pines (eds.). 1988. *The economy as an evolving complex system*. Redwood City: Addison Wesley.

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

Arthur, B., S. Durlauf, and D. Lane. 1997. *The economy as an evolving complex system II*. Redding: Addison Wesley.

Axelrod, R., and M. Cohen. 2000. *Harnessing complexity organizational implications of a scientific frontier*. New York: Free Press.

Bankes, S. 2002. Agent-based modeling: A revolution? *Proceedings of the National Academy of Sciences* 99: 7199–7200.

Camerer, C. 2003. *Behavioral game theory: Experiments in strategic interaction*. Princeton: Princeton University Press.

de Marchi, S. 2005. *Computational and mathematical modeling in the social sciences*. Cambridge: Cambridge University Press.

Epstein, J. 2003. *Generative social science*. Princeton: Princeton University Press.

Epstein, J., and R. Axtell. 1996. *Growing artificial societies: Social science from the bottom up*. Cambridge, MA: MIT Press.

Holland, J., and J. Miller. 1991. Artificial agents in economic theory. *American Economic Review: Papers and Proceedings* 81: 365–370.

Judd, K. 1997. Computational economics and economic theory: Complements or substitutes? *Journal of Economic Dynamics and Control* 21: 907–942.

Judd, K., and L. Tesfatsion (eds.). 2005. *Handbook of computational economics volume 2: Agent-based computational economics*. Amsterdam: North-Holland.

Kirman, A. 1997. The economy as an interactive system. In *The economy as a complex evolving system II*, ed. W. Arthur, S. Durlauf, and D. Lane. Reading: Addison Wesley.

Leombruni, R., and M. Richiardi. 2005. Why are economists sceptical about agent- based simulations? *Physica A* 355: 103–109.

Mason, J., and M. Wellman. 2005. Automated markets and trading agents. In *Handbook of computational economics, volume 2: Agent-based computational economics*, ed. L. Tesfatsion and K. Judd. Amsterdam: North-Holland.

Miller, J. 1998. Active nonlinear tests ANTs of complex simulations models. *Management Science* 44: 820–830.

Nowak, M., S. Bonhoeffer, and R. May. 1994. Spatial games and the maintenance of cooperation. *Proceedings of the National Academy of Sciences* 91: 4877–4881.

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

Page, S., and T. Tassier. 2004. On the existence and stability of Groves Ledyard equilibria. *Journal of Public Economic Theory* 62: 311–335.

Palmer, R., W. Arthur, J. Holland, B. Lebaron, and P. Tayler. 1994. Artificial economic life: A simple model of a stock market. *Physica D* 75: 264–274.

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

Riolo, R., R. Axelrod, and M. Cohen. 2001. Evolution of cooperation without reciprocity. *Nature* 414: 441–443.

Tesfatsion, L. 1997. How economists can get A-life. In *The economy as a complex evolving system II*, ed. W. Arthur, S. Durlauf, and D. Lane. Redding: Addison Wesley.

Tesfatsion, L., and K. Judd (eds.). 2006. *Handbook of computational economics, volume 2: Agent-based computational economics*. Amsterdam: North-Holland.

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

Agents of Production

F. Y. Edgeworth

The causes or requisites of production, often called 'agents of production', may be divided into two classes: human action and external nature; commonly distinguished as 'labour', and 'natural agents'. The first category comprises mental as well as muscular exertion; the second, force as well as matter. To the second factor is sometimes applied the term land: in a technical sense, denoting not only the 'brute earth', but also all other physical elements with their properties. But this term is more frequently employed in another classification, according to which the agents of production are divided into three classes – land, labour, and capital. Of the two classifications which have been stated the former appears the more fundamental and philosophical. That 'all production is the result of two and only two elementary agents of production, nature and labour,' is particularly well argued by Böhm-Bawerk in his Kapital und Kapitalzins, pt. ii. p. 83. 'There is no room for a third elementary source,' he maintains. This view is countenanced by high authorities, of whom some are cited below. Even J.S. Mill, who is disposed to make capital nearly as important as the other members of the tripartite division, yet admits that 'labour and natural agents' are 'the primary and universal requisites of production' (Political Economy, bk. i, ch. iv, §; 1). Prof. Marshall, dividing the subject more closely, thinks 'it is perhaps best to say that there are three factors of production, land, labour, and the sacrifice involved in waiting' (Principles of Economics, p. 614, note).

In the case where both labour and natural agents are required, the most frequent and important case, the question may be raised whether nature or man contributes more to the result. According to Quesnay (*Maximes*, p. 331), land is the sole source of riches. According to Adam Smith, in manufactures 'nature does nothing, man does all' (*Wealth of Nations*, bk. ii, ch. v). The better view appears to be that the division of industries into those in which labour does most and those in which nature does most is not significant. It is like attempting 'to decide which half of a pair of scissors has most to do in the act of cutting' (Mill, *Political Economy*, bk. i, ch. i, § 3).

Agents of production may be subdivided into those which are limited, and those which are practically unlimited. This distinction applies principally to natural agents. For labour may in general be regarded as an article of which the supply is limited. The ownership or use of those agents of production which are limited and capable of being appropriated acquires a value in exchange. Hence rent of land and wages of labour take their origin.

To account for the difference in the rents paid for different lands, it has been usual, after Ricardo, to arrange the lands in a sort of scale of fertility: No. 1, No. 2, and so on. Upon this classification it is to be remarked that productivity, the real basis of the differences in question, does not vary according to any one attribute, such as the indestructible powers of the soil, or proximity to the centres of industry; but upon a number of attributes (compare B. Price, *Practical Political Economics*, chapter on 'Rent'). Moreover, a scale in which lands, or other natural agents, were arranged according to their productive power, would hold good only so long as the other factor of production, human action, might remain constant. A light sandy soil may be more productive than a heavy clay, so long as the doses of labour applied to each are small. But the order of fertility may be reversed when the cultivation is higher. As Prof. Sidgwick remarks 'these material advantages' [afforded by natural agents] 'do not remain the same in all stages of industrial development: but vary with the varying amounts of labour applied, and the varying efficiency of instruments and processes' (*Political Economy*, bk. i, ch. iv, § 3). Compare Prof. Marshall, *Principles of Economics*, bk. iv, ch. iii, § 4.

A similar difficulty attends the attempt to arrange the other agent of production, human labour, in a scale of excellence; whereby to determine what has been called Rent of Ability. Prof. Macvane has noticed this difficulty in an article on 'Business Profits' in the *Quarterly Journal of Economics* (Harvard) for October 1887. Prof.

Walker, in a reply to Prof. Macvane in the same journal, April 1888, admits and very happily illustrates the difficulty (p. 227).

[On this subject as many references might be given as there are treatises on political economy. The twofold classification above indicated is illustrated by the following: Hobbes, *Leviathan*, beginning of ch. xxiv ('The plenty of matter' consists of 'those commodities which from the two breasts of our common mother, land and sea, God usually either freely giveth, or for labour selleth to mankind'). Petty, *Treatises on Taxes* (3rd edn, 1685), ch. viii, p. 57 (labour the father, land the mother, of wealth). Berkeley, *Querist*, Query 4 ('Whether the four elements and man's labour therein be not the true source of wealth'). Cantillon, *Essay*, pt. i, ch. i (land the matter and labour the form of riches). Courcelle-Seneuil, *Traité théorique*, bk. i, ch. iii. Hearn, *Plutology*, ch. ii.

See Also

► Capital as a Factor of Production

Bibliography

Berkeley, G. 1735–37. *The querist*. Dublin.
Böhm-Bawerk, E. 1884. *Kapital und Kapitalzins*. Innsbruck.
Cantillon, R. 1752. *Essai sur la nature du commerce en général*. Paris.
Courcelle-Seneuil, J.G. 1858. *Traité theorique et practique d'économie politique*. Paris.
Hearn, W.E. 1864. *Plutology, or the theory of the efforts to satisfy human wants*. London: Macmillan.
Macvane, S.M. 1887. The theory of business profits. *Quarterly Journal of Economics* 2: 1–36.
Macvane, S.M. 1888. Business profits and wagers. *Quarterly Journal of Economics* 2: 453–468.
Marshall, A. 1890. *Principles of economics*. London: Macmillan.
Mill, J.S. 1848. *Principles of political economy*, 2 vols. London.
Petty, W. 1685. *Treatise on taxes and contributions*. London.
Price, B. 1878. *Chapters on practical political economics*. London.
Quesnay, F. 1846. *Maximes*, ed. E. Daire. Paris.
Sidgwick, H. 1883. *Principles of political economy*. London: Macmillan.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. London.

Aggregate Demand and Supply Analysis

A

Hugh Rose

Temporary Equilibrium

Postulate an elementary period or instant, which may be arbitrarily short. There is a set of parameters given or determined at its outset. They change only from one instant to the next. Within an instant some markets are cleared. In this temporary equilibrium the economy moves from instant to instant in accordance with the laws governing the behaviour of the parameters.

Hicks (1939, p. 122) stated that there will nearly always be some goods whose production can be changed within the instant. Applying this principle to macroeconomics Hicks (1937) treated labour as a perfectly variable factor for the individual entrepreneur, so that, in his interpretation, the Keynesian IS–LM equilibrium, or its full-employment counterpart, is the economy's temporary equilibrium, with employment, output and interest rates determined within the instant, given the parametric stock of capital etc. This is still the standard temporary-equilibrium concept in macroeconomics. A point not lying on the IS curve is usually regarded as indicating a net excess demand for goods.

But there are two serious difficulties. First there is the well-known crux concerning Walras' Law when there is involuntary unemployment in the IS–LM equilibrium. How can there be an excess supply of labour when there is no excess demand for anything else? The ingenious distinction made by Clower (1965, ch. 5) between 'notional' and 'effective' excess demands solves the problem formally, but prompts the question why it is required in macroeconomics when the rest of economics manages without it.

Secondly there is a strong case for assuming that labour, like capital, is a quasi-fixed factor for

the individual entrepreneur, given or determined parametrically at the outset of each instant. For there are costs of hiring and firing people, and even of varying significantly hours worked, at short notice. But this suggests that macroeconomics should be based not on Hicks's principle but on the Marshallian concept of a temporary equilibrium relative to a given state of expectations, in which market prices equate demands in each instant to outputs predetermined at its outset.

Actually a macroeconomic temporary equilibrium of this kind was devised long ago. One of its inventors was Keynes himself. Keynes's economics was Marshallian in this respect from the *Treatise on Money* (1930, chs. 9–11) to the *General Theory* (1936) and beyond. The contrary belief regarding the *General Theory* expressed, for example, by Hicks (1965, pp. 64–6) will be shown to be incompatible with the evidence.

Keynes's object in the *Treatise on Money* (1930, Preface, p. v) was to find a method of analysing dynamic processes towards and around a longer-run equilibrium. With the same end in view we shall present a model of temporary equilibrium under assumptions of constant returns to scale and labour-augmenting technical change, in order that a longer-run equilibrium may be one of steady growth. As in the Marshallian theory of relative prices, the dynamics will depend on revisions of short-term expected (or 'normal') prices when the prices of the temporary equilibrium turn out to be different from them. 'Hicksian' dynamics is somewhat pressed to find convincing substitutes for this lag, on which the Marshallian distinction between market and short-term normal prices is based. We shall show how it can be used in constructing a set of dynamic equations that accomplish Keynes's objective in the *Treatise on Money* and enable us to put into a unified framework a great variety of macrodynamic theories.

But it may be useful to begin by expressing our general approach to aggregative analysis. The subject-matter of macro-economics is, we believe, the behaviour of index numbers, of final output, employment, the stock of capital, interest rates, the general price-level, etc. It is foolish to assume that their components are homogeneous, since index numbers are required just because they are not. We also dissent from the idea that there exists a fundamental non-aggregative system with which they should be consistent. The decision to be made is how far to disaggregate, not how to justify departures from this imaginary construct. Our purpose here will be served at the highest level of aggregation.

The index numbers are taken to reflect the average (or representative) behaviour and experience of economic agents. The deviations from the average are not predicted by the model, and so could not be inferred from it even if everyone knew it in detail.

Supply

We assume a closed economy, so that total money income equals the value of final output. Real final output $Y = Kf(x)$, where K is the inherited stock of capital, $x = N/K$, and N is the demand for labour in efficiency units. For simplicity perfect competition is assumed. At the outset of an instant firms choose x by maximizing the profits expected to accrue in it. Thus optimum x depends on *short-term expectations*. If p is an index of prices expected for the instant and w an index of money wages per efficiency unit of labour, x maximizes $pf(x) - wx$. Necessary and sufficient conditions for an interior maximum are $f'(x) = w/p$ and $f'(x) < 0$ Given p, w, and K, then, both output, Y, and the sum of expected money incomes, pY, are parameters for the instant.

Prices and Windfall Profits

Actual incomes may differ from pY. Let Q be the net sum of unexpected incomes deflated by pK. Thus money incomes deflated by K are $p[f(x) + Q]$. If π is the price level of final output, by definition $\pi f(x) = p\ ([f(x) + Q]$, so that Q will turn out to be $\gtreqless 0$ according as the market determines π to be $\gtreqless p$ within the instant. Since output is completely inelastic within the instant, $pQ = (\pi - p)\ f(x) = [\pi f(x) - wx] - [pf(x) - wx]$ is the net sum of unexpected or *windfall* profits deflated by K. (In the *Treatise on Money* Keynes apparently defined windfall profits as the excess of entrepreneurs' actual over *long-term normal* renumeration (1930, pp. 124–5). The definition here follows from our having adopted his assumption in the

General Theory (1936, ch. 5) that current employment of labour depends on *short-term* expectations, so that windfalls become the excess of actual over *short-term expected* profits.)

Excess Demand for Final Output

We assume pK-deflated planned investment and saving to be functions $I(Q, r, x)$ and $S(Q, r, x)$ Planned saving is expected income minus planned consumption. r is an index of the general level of real interest rates. The pK-deflated excess demand for final output is

$$X_g = I(Q, r; x) - S(Q, r; x) - Q$$

The subscript g is for 'goods'. The semicolon preceding x indicates that it is a parameter for the instant. I_Q may be negative, since unexpectedly high prices may induce disinvestment in inventories. The sign of S_Q is ambiguous: a negative income effect may be outweighed by a positive substitution effect of unusually high or low π in relation to p. So is the sign of S_r. But we assume that $I_Q - S_Q - 1$ and $I_\lambda - S_r$ are both negative. I_x is non-negative, but is positive if *long-term expectations* of profit move in the same direction as short-term expectations of it. Finally S_x, which has the sign of the marginal propensity to save, is assumed to be positive.

π will rise or fall (given p) according as X_g is positive or negative. So, therefore, will Q.

Excess Flow Demand for Money

There is a central-banking system. We abstract from the note issue. Commercial banks' reserves at the Central Bank, deflated by pK, are R. The public's pK-deflated demand for commercial banks' deposits is a function, $L(Q, r; x, \lambda)$ where λ is the parametric expected rate of inflation of p. L_r is negative and so is L_λ. L_Q may be zero. In any case its sign is ambiguous. There may be a positive income effect. But since a portion of loans is normally kept on deposit, when a rise in Q reduces the demand for inventories (and correspondingly the demand for bank loans) the borrowers' demand for deposits may also be reduced. Finally L_x may be negative. For the rise in expected profits with x may increase confidence, reducing the demand for liquidity. (Transactions demand is already largely accounted for by expressing the demand as a ratio to pK.)

The public as a whole can make deposits whatever it wishes them to be by altering its borrowings from the banks. There can be no inevitable net creation of 'derivative' deposits by the banks themselves as they attempt to remove a net surplus of reserves, when the public commands the volume of bank loans at the banks' current loan rates. For a discussion of the genesis of deposits see Rose (1985, section 4).

It is convenient, but not essential, to assume that deposits are momently equal to the stock demand for them. The banks, however, have, at the outset of an instant, reserves that are not, in general, what they need. Let c be their desired ratio of reserves to deposits, assumed constant for simplicity. When cL differs from R they try to reduce the gap during the instant by *active net hoarding*. Its extent is assumed to be ß$(cL - R)$, where ß is a positive adjustment coefficient.

But there may also be *passive* net hoarding. The theory of the precautionary demand for money suggests that, since the terms for unexpected transactions between money and securities at short notice are apt to be worse than those for expected transactions between them, the optimum strategy should involve a temporarily passive response to unexpected net receipts, i.e., passive net hoarding of them. Now unexpected net receipts arise when Q is non-zero. We therefore assume that passive net hoarding is $\alpha Q (0 \leq \alpha \leq 1)$, with α constant.

The pK-deflated excess flow demand for money (reserves and deposits) is therefore

$$X_m = \beta[cL(Q, r; x, \lambda) - R] + \alpha Q - \dot{R}$$

the subscript m is for 'money'.

Walras' Law

Since final output is a parameter, the temporary equilibrium is an equilibrium of exchange. The sum of the values of the excess demands for goods, securities, and money must be zero. The excess demands for factors are irrelevant during the instant, owing to the assumption that factor employments are fixed at its outset. The problem

encountered in the Hicksian theory simply does not arise here.

Excess Flow Demand for Loanable Funds

The excess supply of securities is the excess demand for lonable funds, whose pK-deflated value is X_f. Therefore by Walras' Law

$$X_f = I(Q, r; x) - S(Q, r; x) - Q + \beta[cL(Q, r; x, \lambda) - R] + \alpha Q - \dot{R}$$

The subscript f is for 'funds'. r will rise or fall according as X_f is positive or negative.

The Temporary Equilibrium with Parametric R

If the Central Bank sets R for the instant, $\dot{R} = 0$. The adjustment of r and π (or equivalently Q) puts X_f and X_g to zero, establishing a unique equilibrium if, in addition to the inequalities $I_Q - S_Q - 1 < 0$, $I_r - S_r < 0$, and $L_r < 0$, the condition $(I_Q - S_Q - 1)L_r - (I_r - S_r)L_Q > 0$ is satisfied.

The equations are

$$I(Q^*, r^*; x) - S(Q^*, r^*; x) = Q^*\alpha[I(Q^*, r^*; x) - S(Q^*, r^*; x)] = \beta[R - cL(Q^*, r^*; x, \lambda)]$$

$$\pi^* = p + p(I^* - S^*)$$

(The asterisks indicate equilibrium values.) The first is Keynes's Fundamental Equation (viii) (1930, p. 138). The third is the form assumed by his Fundamental Equation (iv) (1930, p. 137) when windfalls are defined as in section "Prices and Windfall Profits" above. The second is more general than its counterpart in Keynes. For he assumed that there is no passive net hoarding, i.e., that $\alpha = 0$. The consequence is his 'liquidity preference' theory of interest, $L(Q^*, r^*; x, \lambda) = R/c$.

He held to this aspect of his temporary equilibrium not only in the *Treatise on Money* and immediately after it (Keynes 1973a, pp. 224–5) but also in and after the *General Theory*. The net demand for funds represented by I^*-S^* is matched by net loans from windfalls exactly equal to it. Thus in a letter to Hawtrey written soon after the publication of the *General Theory* he insisted that an increase in investment would not directly raise r^* because it would raise the demand for securities by precisely the same amount (Keynes 1973b, p. 12).

But if α is positive I^*-S^* is not fully matched by net loans from windfalls. Active net dishoarding must fill the gap, viz. αI^*-S^*, and r^* must stand above or below the level corresponding to $L^* = R/c$ according as I^*-S^* is positive or negative. This is essentially the 'loanable funds' theory of interest, for which see, e.g., Robertson (1940, pp. 1–20).

The Question of Say's Law

If rational conduct does imply that α is positive, there is a decisive answer to the question whether aggregate demand must be a determinant of the economy's behaviour, or equivalently whether the 'classical' theory of interest (Keynes 1936, ch. 14) must be wrong. (For a fuller account of this subject see Rose (1985, pp. 1–17).) If for each pair of values of x and λ we can find a stock of reserves with which the temporary-equilibrium equations become

$$I(0, r^*; x) = S(0, r^*; x)$$

$$cL(0, r^*; x, \lambda) = R^*$$

$$\pi^* = p$$

with $r^* > 0$, the answer is no. The 'classical' theory of interest becomes valid, and, since $Q^* = 0$, aggregate money demand, $p[Y + K(I^* - S^*)]$, and money income, $\pi^* Y$, are equal to and determined by the given sum of expected incomes, pY. If such an R^* could always be found, inflation, fluctuations, unemployment there might be, but none of them due to movements of aggregate demand for output. Moreover the appropriate level of reserves can be found and sustained 'without the necessity for any special intervention or grandmotherly care on the part of the monetary authorities' (Keynes 1936, p. 177). In effect Say's Law of Markets can be imposed whenever we wish; for the market mechanism itself will guarantee that supply, pY, creates its own demand. To impose it the Central Bank should stand passively

ready to deal in securities with the member banks, at their current prices, in exchange for reserves. Both convenience and economic incentive will induce the banks to accomplish their active net hoarding via the Central Bank, the incentive being the tendency of security prices to move against them if they go to the market instead. Thus they will adjust their reserves to the "demand" for them in accordance with the equation $\dot{R} = \beta(cL - R)$. But then, since α is positive, the second equation in section "The Temporary Equilibrium with Parametric R" above implies $I^* = S^*$. The market cannot support a non-zero I^*-S^* when the banks provide it with no active net dishoarding.

But if α were zero the second equation in section "The Temporary Equilibrium with Parametric R" would not imply $Q^* = 0$ when $R^* = cI^*$. Instead there would be many possible equilibria. Which of them would eventuate would depend on which value of R^* were fortuitously reached in the adjustment to cL^*. The Central Bank's policy could not succeed in imposing Say's Law. It would simply render indeterminate the equilibrium at which I^*-S^* was matched by net loans from windfalls. No wonder Keynes was so insistent on his 'liquidity preference' theory of interest!

In a system with no Central Bank, all money consisting of the notes and deposits of *non-colluding* commercial banks holding each others' deposits as reserves, Say's Law would always rule if α were positive. For if $R = c = 0$ then $Q^* = 0$.

Process Analysis

Comparative Statics of Temporary Equilibrium

Let m be the 'potential' supply of deposits, R/c. We shall refer to it as the supply of money deflated by pK.

The temporary equilibrium implies functions $Q^*(x, m, \lambda)$ and $r^*(x, m, \lambda)$. The signs of their partial derivatives are of the first importance in process analysis. What can be learnt about them from the formulae obtained by differentiating the equations of section "The Temporary Equilibrium with Parametric R" and applying Cramer's Rule, together with the inequalities assumed there?

Definite signs are attached to r^*_m, r^*_λ, Q^*_m, and Q^*_λ. The first two are negative, of course, and in consequence the last two are positive.

Sign $Q^*_x = \text{sign}[(I_r - S_r)L_x - (I_x - S_x)L_r]$. It may easily be positive; for the marginal inducement to invest, I_x, may exceed the marginal propensity to save, S_x, and L_x may be negative (see section "Excess Flow Demand for Money").

Sign $r^*_x = \text{sign}\left[(\alpha + \beta c L_Q)(I_x - S_x) - (I_Q - S_Q - 1)\beta c L_x\right]$. If L_Q is zero and if, as one might expect, β is large, sign $r^*_x = \text{sign} L_x$.

Since the banks' desired cash ratio will make no further explicit appearance, the letter c will be given a new definition in section "The Equations of Motion" below.

Capital Accumulation

Since the goods markets are cleared, actual and planned investment are equal, Therefore $\dot{K}/K = I^*(x, m, \lambda)$. An essential requirement i snome theories of growth and all 'over-investment' theories of the business cycle (Haberler 1937, ch. 3) is that I^*_x should be non-negative. Now $I^*_x = I_x + I_r r^*_x + I_Q Q^*_x$, so that all is well if I_r and L_x are negative and $|I_Q|$ is small. In a Say's-Law regime $I^*_x = (I_x S_r - I_r S_x)/(S_r - I_r)$, which is almost surely positive. The other two partials are positive if I_r is negative and $|I_Q|$ small.

The Dynamics of Short-Term Expectations

Three forces act on p from one instant to the next, namely expected inflation, the excess of windfall profits over windfall losses, and what we may call cost push. Their action is expressed by

$$\dot{p}/p = \lambda + H(Q^*) + \sigma(\dot{w}/w - \lambda), 0 \leq \sigma \leq 1, .$$

with σ constant. H is an increasing function and $H(0) = 0$, because windfalls cause trial-and-error revision of short-term expectations. When $Q^* = 0$ and $\dot{w}/w - \lambda$ the inflation of expected prices equals the expected inflation of them, λ. The cost push term, $\sigma(\dot{w}/w - \lambda)$, allows for the possibility that when the index of efficiency wages rises or falls, firms expect prices to rise or fall in other affected industries, diverting demand to or from their own industry.

The Dynamics of Efficiency Wages

Similarly three forces act on w, namely expected inflation, the excess demand for labour, and the indexation of wages to expected prices. Their action is expressed by

$$\dot{w}/w = \lambda + F(x/\upsilon) + \sigma(\dot{p}/p - \lambda), 0 \leq \tau \leq 1,$$

with τ constant. Let N^s be the supply of labour in efficiency units and υ be N^s/K. Then $x/\upsilon = N^s/K$. When $x = \upsilon$ unemployment equals unfilled vacancies. The corresponding unemployment rate is the 'natural rate, kept in being by the break-up of old jobs and imperfect information about the new jobs that replace them. (Firms with vacancies use their workers more intensively while seeking to fill them, so that the vacancies do not preclude the production of Y $Kf(x)$.) The unemployment rate is a decreasing function of x/υ. Unemployment is involuntary when x/v is <1. F is a non-decreasing function with $F(1) = 0$.

The Equations of Motion

Logic requires $\sigma\tau<1$; for $\dot{w}/w/\dot{p}/p$ cannot be both exclusively determined by x/υ and exclusively determined by Q^*. Therefore the development of the economy is governed by the following equations:

$$\begin{aligned}\dot{x}/x &= aH[Q^*(x,m,\lambda)] - bF(x/v)\dot{p}/p - \lambda \\ &= cH[Q^*(x,m,\lambda)] - gF(x/v)\dot{w}/w - \lambda \\ &= hH[Q^*(x,m,\lambda)] - cF(x/v)\dot{v}/v \\ &= n - I^*(x,m,\lambda).\end{aligned}$$

The first is from the derivative of log $f'(x) = \log w/p$ with respect to time. The second and third combine the equations of sections "The Dynamics of Short-Term Expectations" and "The Dynamics of Efficiency Wages." The fourth is from $\dot{v}/v = \dot{N}^s/N^s - \dot{K}/K$, with n defined as N^s/N^s, the growth of the supply of labour in efficiency units. The coefficients are as follows:

$$\begin{gathered}a = \phi(1-\tau)/(1-\sigma\tau) \geq 0 \quad \text{with}\, \phi = -f'(x)/xf''(x) > 0; \\ b = \phi(1-\sigma)/(1-\sigma\tau) \geq 0; \quad c = 1/(1-\sigma\tau) > 0; \\ g = \sigma/(1-\sigma\tau) \geq 0; \quad h = \tau/(1-\sigma\tau) \geq 0.\end{gathered}$$

In conjunction with particular assumptions about the behaviour of m, λ, and n, these equations enable us to capture the essential characteristics of many macrodynamic theories and to display their interrelationships.

Processes with a Constant Labour–Capital Ratio

If $v = N^s/K$ is a constant, $\bar{v}$, the last equation in section "The Equations of Motion" disappears. Two interpretations are possible: either the change in υ over the relevant period is negligible, or labour-augmenting technical progress equals the growth of capital per worker. Processes with constant υ can therefore be regarded as occurring in relation either to a short-period equilibrium without technical change or to a long-period equilibrium with endogenous growth. The formal structure is the same in both cases.

Keynes's General Theory

Expectations and Short-Period Equilibrium

In the *General Theory* the temporary equilibrium converges to a Marshallian short-period equilibrium with no technical change. Keynes imagines two ways by which it may be reached. In the *General Theory* for the most part he assumes as a short cut that short-term expectations are always fulfilled (Keynes 1973a, pp. 602–3). At the outset of an instant, entrepreneurs, correctly anticipating the aggregate demand-price, choose the employment, x, that will maximize their actual profit, $\pi^* f(x) - wx$, since $p = \pi$. This is the case of the 'instantaneous multiplier'; Y^* is determined at the outset of each instant so as to make $Q^* = 0$, i.e., $I^* = S^*$, within it. However he does not insist on this. If short-term expectations are not always fulfilled, p is adjusted by trial and error from one instant to the next. This process, along with the assumption that during it the economy is in the temporary equilibrium, is actually contemplated at one point in the *General Theory* itself (Keynes 1936, pp. 123–4), and indeed later he wished that he had made more of it there (Keynes 1973b,

pp. 180–1). We may also wish he had; for by not doing so he originated the myth that he was himself rejecting the *Treatise on Money's* Marshallian conception of temporary equilibrium in favour of Hick's conception of it.

Money Wages and Employment

Keynes claims as a fundamental objection to the 'classical' theory the postulate that the real wages, w/p, on which employment, x, depends are *directly* affected by labour's bargaining about money wages (Keynes 1936, p. 13). Keynesian unemployment is involuntary in a special way: it cannot be directly eliminated by flexibility of money wages. This dogma is first enunciated in the *Treatise on Money* (Keynes 1930, p. 167), where changes in w have no direct tendency to bring about non-zero profits, Q^*, because, so long as they are not allowed to affect interest rates, they cause a proportionate change in the price level, $\pi^* = p + p\ (I^* - S^*)\ f(x)$. But that is so only if they induce a proportionate change in expected prices, p, leaving w/p, and so x, unaffected. In fact he is assuming full cost push, $\sigma = 1$, so that, in section "The Equations of Motion" $\dot{x}/x = aH(Q^*)$. Changes in employment are due solely to the effect of Q^* on *short-term expectations of prices in terms of wage units*, p/w, not at all to changes in the wage unit, w, itself, except in so far as they may affect the parameters determining Q^*.

Not a strong foundation for a *general* theory! Nevertheless there is a good reason for retaining this possibility in our process analysis. In the *Hicksian* temporary equilibrium the real wage is likewise determined independently of the money wage so long as m is given. When post-Keynesians who adopt the Hicksian viewpoint allow for some degree of money-wage flexibility, the qualitative behaviour of their models will be just as if there were full cost push.

The Trial-and-Error Process

If, for simplicity, one treats as a parameter the supply of money 'in terms of wage units', so that $m = kf'(x)$ with k a positive constant, the process, with parametric λ, is $\dot{x}/x = aH\ [Q^*(x, kf'(x); \lambda)]$, $\dot{p}/p - \lambda = c\ [H\ (Q^*) + F\ (x/\overline{v}\,)]$, $\dot{w}/w - \lambda = cF\ (x/\overline{v}\,)$; for $g = c$ when $\sigma = 1$, and $h = 0$ because no wage-indexation is assumed. In the equilibrium [which is stable if m $Q_x^* + Q_m^* k f''(x)$ is negative] $Q^* = 0$, $x^* < \overline{v}$, and $(\dot{p}/p)^* - \lambda = (\dot{w}/w)^* - \lambda = F\ (x^*/\overline{v}\,)$. Thus Keynes really needs to assume wage inflexibility below full employment, $F\ (x/\overline{v}\,) = 0$ for $x < \overline{v}$, in addition to $\sigma = 1$. Otherwise the equilibrium would be upset by a systematic error about expected inflation. The underemployment equilibrium is then $I\,(0, r^*, x^*) = S\,(0, r^*, x^*)$, $L\,(0, r^*, x^*; \lambda) = m^* = kf'(x^*)$, $(\dot{p}/p)^* = (\dot{w}/w)^* = \lambda$, with $x^* < \overline{v}$.

Say's Law

If m^* is such that $Q^* = 0$ for all x and λ, the process is $\dot{x}/x = -bF\ (x/\overline{v}\,)$, $\dot{p}/p - \lambda = gF\ (x/\overline{v}\,)$, $\dot{w}/w - \lambda = cF\ (x/\overline{v}\,)$. When σ is less than unity and F is strictly increasing there is a convergence to equilibrium at the natural unemployment rate, with inflation of p and w at the rate λ, which is not determined by the system. The equilibrium equations are $I\ (0, r^*, x^*) = S\ (0, r^*, x^*)$, $x^* = \overline{v}$, $L\ (0, r^*, x^*; \lambda) = m^*$, $(\dot{p}/p)^* = (\dot{w}/w)^* = \lambda$.

Keynes (1936, p. 26) maintained that Say's Law would imply indeterminacy of x. Indeed it would under his assumption $\sigma = 1$, for then $b = 0$. However his allegation, that in these circumstances competition between entrepreneurs would lead to full employment, is a nonsequitur, as Hawtrey pointed out to him (Keynes 1973b, pp. 31–2).

Full Wage-Indexation

If $\tau = 1$ then $a = 0$. The process is $\dot{x}/x = -bF\ (x/\overline{v}\,)$, $\dot{p}/p - \lambda = cH[Q^*\ (x,m,\lambda)] + gF\ (x/\overline{v}\,)$, $\dot{w}/w - \lambda = c[H\ (Q^*) + F\ (x/\overline{v}\,)]$. As under Say's Law, there is convergence to the natural unemployment rate. But, whereas Say's Law leaves inflation indeterminate, full wage-indexation offers a painless means of manipulating it by changing the supply of money.

Underemployment Equilibrium in a Growing Economy

The Keynesian equilibrium of section "The Trial-and-Error Process" can be interpreted as one of endogenous growth with involuntary unemployment. This extension is due to Domar (1946, pp. 137–47). Actually he used the 'extreme

Keynesian' assumptions that I is determined by entrepreneurs' animal spirits, with $I_x = I_r = 0$, and that $S = sf(x)$ with s a positive fraction, so that money has no effect on them. In his equilibrium (which is obviously stable, given I) $I = sf(x^*)$, and the ratio of actual output, $Y^* = Kf(x^*)$, to normal capacity output, $P = Kf(\overline{\upsilon})$, is less than unity unless I is large enough to imply $x^* = \overline{\upsilon}$.

Business Cycles with a Constant Labour-Capital Ratio

A Purely Monetary Theory of Cycles

The appellation is taken from Haberler (1937, ch. 2), where he expounds Hawtrey's theory, contrasting it with overinvestment theories, in which changes in υ are an essential feature. The following version generalizes a model constructed by Phillips (1961, pp. 360–70) but conveying ideas much like those expressed by Hawtrey. For his first statement of them see Hawtrey (1928, ch. 5).

There are four assumptions: (i) F is strictly increasing; (ii) the ratio of the nominal money-suppy to K grows at the constant rate μ; (iii) people expect inflation to be μ, i.e., $\lambda = \mu$; (iv) the equilibrium is stable.

Since m is the supply of money deflated by pK, $\dot{m}/m = \mu - \dot{p}/p$ by (ii). But from the second equation of motion in section "The Equations of Motion" we have $\dot{p}/p = \lambda + cH + gF$, so that $\dot{m}/m = \mu - \lambda - cH - gF = -cH–gF$ by (iii). Hence the dynamic system in x and m is

$$\begin{aligned}\dot{x}/x &= aH[Q^*(x,m;\mu)] - bF(x/\overline{\upsilon})\dot{m}/m \\ &= -cH[Q^*(x,m;\mu)] - gF(x/\overline{\upsilon}),\end{aligned}$$

with the equilibrium $x^* = \overline{\upsilon}, I(0, r^*, x^*) = S(0, r^*, x^*), L(0, r^*, x^*; \mu) = m^*$. Notice that changes in μ have no real effect on it, merely altering m^*.

There is local stability if $\overline{\upsilon}\left[aH'(0)Q_x^* - bF_x\right] - m^*cH'(0)Q_m^*$ is negative. Thus even if the first term, representing the effect of x on $\dot{x}$, is positive, the second term, representing the effect of p on the course of real balances, and therefore on the course of interest rates, can (and we are assuming will) outweigh it. For Hawtrey the first term is positive. A shock induces a cumulative expansion (or contraction), which is eventually reversed because a growing shortage (or abundance) of money increases (or reduces) interest rates.

The discriminant of the linearized system is

$$\begin{aligned}D = &\left\{\overline{\upsilon}\left[aH'(0)Q_x^* - bF_x\right] - m^*cH'(0)Q_m^*\right\}^2 \\ &- 4(ag + bc)\overline{\upsilon}m^*H'(0)F_xQ_m^*.\end{aligned}$$

It implies that there will be oscillations if, *ceteris paribus*, Q_m^* is large. For $\partial D/\partial Q_m^*$ is negative when the stability condition is satisfied.

Examination of D reveals a very interesting point. With full cost push ($b = 0$) higher wage-flexibility (larger F_x) must, *ceteris paribus*, induce more rapid oscillations. Compare Keynes (1936, pp. 269–71). (Phillip's model, in which a coefficient β corresponds with our F_x, has this Keynesian characteristic.) For it increases the frequency of the turning points induced by the monetary factor without damping the cumulative process. But when b is positive high enough wage-flexibility eliminates the cumulative process entirely. No oscillations can occur.

Staglation Cycles

There have been periods during which inflation and the unemployment rate have risen or fallen simultaneously. Three assumptions are sufficient to explain this phenomenon: (i) expectations of inflation are adaptive: $\lambda = \gamma(\dot{p}/p - \lambda)$ with γ positive and constant; (ii) monetary policy is to decrease (or increase) m when λ rises (or falls): $m = m(\lambda;\theta)$ with m_λ negative; θ is a shift parameter with m_θ positive; (iii) the equilibrium is stable.

We have then

$$\dot{x}/x = aH[Z(x,\lambda;\theta)] - bF(x/\overline{\upsilon})$$

$$\dot{\lambda} = \gamma\{cH[Z(x,\lambda;\theta)] + gF(x/\overline{\upsilon})\},$$

where $Z(x, \lambda;\theta)$ is $Q^*[x, m(\lambda;\theta), \lambda]$. The equilibrium equations are $x^* = \overline{\upsilon}, I(0, r^*, x^*) = S(0, r^*, x^*)$, and $L(0, r^*, x^*, \lambda^*) = m(\lambda^*;\theta)$. Observe that changes in θ affect only λ^*.

The equilibrium is locally stable if Z_λ and $\overline{\upsilon}[aH'(0)\ Z_x - bF_x] + \gamma\ cH'(0)\ Z_\lambda$ are both

negative. The first condition is satisfied if and only if m_λ is more negative than L_λ.The authorities must ensure that real interest rates move in the same direction as λ The second condition guarantees that the course of real interest rates eventually dissipates the cumulative expansions and contractions that may occur if $Z_x = Q_m^*$ is large.

If there are oscillations the turning points are due to the Central Bank's policy. As in the previous model higher wage-flexibility increases their frequency if b is zero, but weakens the cumulative forces if b is positive.

A shock due to a change in θ must initially cause x and λ to move in the same direction. But, whereas λ tends to a new equilibrium, x must tend back to the original $x^* = \overline{\upsilon}$. There must therefore be a period during which x and λ move in opposite directions, and since the inflation of both expected and actual prices tends to λ^*, there must also be a period during which inflation and the unemployment rate move in the same direction.

Keynesian Overinvestment Cycles

Henceforward we assume that $\dot{N}^s/N = n$ is a constant, thereby resuscitating the fourth equation in section "The Equations of Motion."

Purely monetary theories fail to reproduce two observed features of business cycles: (1) The unemployment rate continues to fall (or rise) after entrepreneurs' expected profit-rates have begun to fall (or rise). (2) The real efficiency wage is not a monotonically increasing function of the unemployment rate. But overinvestment theories with a variable unemployment rate do reproduce them.

Natural and Warranted Rates of Growth

The natural rate is n, the sum of the growth rates of the supply of workers and efficiency per worker. The term 'warranted rate' was introduced by Harrod (1939, pp. 14–33) to designate a rate of growth of output which, if it occurs, will leave all parties satisfied that they have produced the right amount (ibid., p. 16). Several formulae are given for it there, and also in Harrod (1948, Lecture 3) and Harrod (1952, Essay 14), depending on alternative assumptions about the determinants of planned investment and planned saving. But the alternatives have one thing in common, namely that these plans are not significantly influenced by monetary policy; either the real rate of interest cannot easily be changed, or the plans are inelastic with respect to it (Harrod 1952, pp. 95–100). Theories involving the warranted rate have an 'extreme Keynesian' bias.

In our equations of motion assume (i) $b = 0$; (ii) F is zero on a large interval around $x = \upsilon$; (iii) Q^* and I^* depend only on x; (iv) $\tau = 0$. Then $\dot{x}/x = aH[Q^*(x)]$, $\dot{p}/p - \lambda = cH[Q^*(x)]$. $\dot{w}/w = \lambda$, and $\dot{\upsilon}/\upsilon = n - I^*(x)$. The warranted rate is $\dot{Y}/Y = I^*$with $Q^* = 0$, for it is justified by the realization, on the average, of short-term expectations.

Now as it stands this system is quite useless. The warranted rate is divorced from the natural rate, so that there is almost surely no equilibrium. But the defect can be remedied if either I or S can be assumed to depend on υ.

Autonomous Consumption

A rationale for making S depend on υ was given by Matthews (1955, pp. 75–95), who suggested that planned consumption from a given income increases with the unemployment rate. Support for the unemployed is at the expense of planned saving. Such changes in consumption are 'autonomous' in that they are not in response to changes in income. Thus $S^* = S^*(x,\upsilon)$, with S_υ^* negative. The system

$$\dot{x}/x = aH[Q^*(x,\upsilon)]$$

$$\dot{\upsilon}/\upsilon = n - I^*(x)$$

is assumed to have a unique equilibrium, $n = I^*(x^*) = S^*(x^*,\upsilon^*)$, with underemployment, i.e., $x^* < v^*$.

Shock-Induced Oscillations

Assume that the equilibrium is stable. This is the case if $Q_x^* = I_x^* - S_x^*$ is negative and I_x^* is positive

at the equilibrium point. It can be shown that there will be oscillations if $|Q_x^*|$ is sufficiently small. A shock induces overinvestment cycles, in that during the boom the growth of capital is excessive ($I^* > n$).The upper turning point is reached when the consequential fall in υ pushes S^* above I^*. Similarly the lower turning point is reached when the rise in υ, due to an excess of n over I^* during the slump, pushes S^* below I^*. For this kind of theory see Samuelson (1939, pp. 75–8). In his version n is zero and autonomous consumption spending is by the government.

Self-Exciting Oscillations

Three conditions are sufficient for these: (i) The equilibrium is unstable but I_x^* is positive; (ii) nevertheless Q_x^* is negative for high and low values of x, say because short-term expected profit seems a less trustworthy guide to investment planning when it has moved far from its equilibrium; (iii) $H'(Q^*)$ is so large that the changes in x when Q^* is non-zero are much larger than the changes in υ when I^* differs from n. By (i) the equilibrium is surrounded by centrifugal forces, and is almost surely not the initial state. By (ii) there are turning points for x, because when x and υ are moving in opposite directions they combine to reduce windfalls, $|Q_x^*|$. By (iii) there are turning points for υ, because of the rapidity with which net overinvestment, $|I^* - n|$, is reduced when x and υ are moving in the same direction.

This essentially is Kaldor's theory (Kaldor 1940, pp. 78–92). Only the first two conditions are given in his text, but the third is implicit there, and is explicitly stated in his appendix (Kaldor, p. 90).

Autonomous Investment

Some investment may grow at the natural rate, n. Then $I^* = J(x) + Ae^{nt}/K$, where $KJ(x)$ is 'induced' investment, Ae^{nt} is 'autonomous' investment, and A is a positive constant. Since $e^{nt} = N^S/N_0^S, I^* = J(x) + \left(A/N_0^S\right)\upsilon$, or, more generally, $I^* = I^*(x, \upsilon)$ with I_υ^* positive. The system

$$\dot{x}/x = aH[Q^*(x, \upsilon)]$$

$$\dot{v}/v = n - I^*(x, \upsilon)$$

is assumed to have a unique equilibrium, $n = S^*(x^*) = I^*(x^*, \upsilon^*)$, with $x^* < \upsilon^*$.

Shock-Induced Oscillations

Assume that the equilibrium is stable. This is so if $x^* aH'(0)Q_x^* - \upsilon^* I_x^*$ is negative and $I_x^*(x^*, \upsilon^*)$ is positive. It can be shown that there must be oscillations if, *ceteris paribus*, Q_x^* is large. Overinvestment (underinvestment) leads to an upper (lower) turning point as changes in υ push I^* below (above) S^*. For this alternative to the autonomous-consumption story see Kalecki (1939, Essay 6). He assumes that n is zero.

Self-Exciting Oscillations

A persistent cycle follows from assumptions similar to those of Hicks (1950); cf. also Goodwin (1951, pp. 1–17): (i) The equilibrium is unstable. (ii) There is a full-employment *ceiling*, a rigid x barrier, C, such that $x \leq C\upsilon$.It is a constraint on x that is binding so long as its free motion would violate it. (iii) There is a value of x, viz. $\xi < x^*$, such that $I_x^*(x, \upsilon)$ is positive for all $x > \xi$ but is zero for all $x \leq \xi$.For induced gross investment in fixed capital cannot be negative, and further induced disinvestment in inventories would disrupt the productive process (cf. Hicks 1950, p. 104).

The cycle is attained in finite time from any non-equilibrium initial state. It has a *floor* implied by the fact that, if in its course the situation $I^*(x,\upsilon) = n$ occurs when $x \leq \xi,\upsilon$, must remain constant until x has risen above ξ. The floor value of υ is the solution to $I(\xi,\upsilon) = n$. The cycle must hit either the ceiling or the floor, but need not hit both.

Non-Keynesian Overinvestment Cycles

Henceforth we assume $\sigma < 1$ and some flexibility of money wages.

Oscillations with Imperfect Wage-Flexibility

F is strictly increasing, and there are positive constants q and l ($q > 1 > l$) such that F tends

to $+\infty$ as x/υ tends to q, and to $-\infty$ as x/υ tends to l.

A 'Non-Monetary' Theory

Under a Say's-Law regime

$$\dot{x}/x = -bF(x\upsilon)$$

$$\dot{\upsilon}/\upsilon = n - I^*(x),$$

Where $I_x^* = (I_x S_x - I_r S_x)(S_r - I_r) > 0$ (see section "Capital Accumulation"). The equilibrium, $n = I^*(\upsilon^*)$, is globally stable, but there will be shock-induced oscillations if F' is small and I_x^* is large in its neighbourhood. For the analysis and a comparison with Cassel's theory see Rose (1969, section III).

There will also be such oscillations if the elasticity of substitution between labour and capital (and therefore b) is small. The model then reproduces approximately Goodwin's growth cycle (Goodwin 1967, pp. 54–8). (If, as he assumes, the elasticity is zero, and in addition all profits are saved and all wages consumed, every solution will be periodic in w/p and υ).

If, however, wages were perfectly flexible the system would reduce to $\dot{x}/x = n - I^*(x)$ which is Solow's growth model (Solow 1956, pp. 58–94).

A Monetary Theory

Let monetary policy be to sustain a constant m. The system

$$\dot{x}/x = aH[Q^*(x; m, \lambda)] - bF(x/\upsilon),$$

$$\dot{\upsilon}/\upsilon = n - I^*(x; m, \lambda)$$

$$\dot{p}/p - \lambda = cH + gF$$

has only a 'quasi-equilibrium' if λ is arbitrarily given: for $\dot{x} = \dot{\upsilon} = 0$ does not imply $\dot{p}/p = \lambda$. To avoid this systematic error about long-run inflation we assume that the public foresees the value λ must take if $(\dot{p}/p)^*$ is to equal it. The equilibrium will then be $n = I(0, r^*, x^*) = S(0, r^*, x^*)$, $L(0, r^*, x^*, \lambda^*) = m$, $v^* = x^*$.

The interesting characteristic of this model is that, if the equilibrium is unstable and if I_x^* is everywhere positive, there must be self-exciting oscillations whose amplitude can be quite small. For the details see Rose (1967, pp. 153–73).

An Equilibrium Theory of Business Cycles

Once upon a time cycles were thought to arise from unsustainable alternations in the structure of the production, brought about by inappropriate and unanticipated changes in the supply of money. Wage inflexibility was not an essential ingredient. This position, held by Hayek (1935, Lecture III) and a cohort of 'Austrian' economists, is surveyed in Haberler (1937, pp. 31–67). Recently, *Lucate duce*, there has been a remarkable attempt to recapture it (Lucas 1975, pp. 1113–44).

The assumptions in our version of it are as follows: (i) there is continuous full employment; (ii) the growth rate of nominal money per unit of capital is a constant, μ. (iii) $\lambda = \mu$ (iv) there is no cost push ($\sigma = 0$). Therefore

$$\dot{x}/x = n - I^*(x, m; \mu)$$

$$\dot{m}/m = -H[Q^*(x, m; \mu)]$$

The equilibrium is almost certainly stable, but there can be oscillations if I_x^* and Q_m^* are small and Q_x^* is negative.

For simplicity we tell the story as if $n = \mu = \lambda = 0$. Equilibrium is disturbed by an unanticipated increase in nominal money. Interest rates fall, creating an investment boom and net windfall profits ('forced saving'). The investment boom increases capital, output, and capital intensity, $K/Y = 1/f(x)$, and is only weakly checked by the larger capital (lower x). But net windfalls raise p (reduce m) and so interest rates rise, eventually leading to an upper turning point for K and K/Y. Net windfalls are still positive, but, once K begins to fall, both higher interest rates and lower K (higher x) convert them into net losses. Now both K is falling and there are net windfall losses. But these reduce p and so interest rates fall, leading to a lower turning point for K and K/Y. Finally lower interest rates and higher K create net windfall profits once again, and a new boom of investment and windfalls begins.

This version may not please Lucas and his school. Persistent, recurrent, and unexploited profit

opportunities are anathema to them. But for the inhabitants of their archipelago there persist also recurrent, unexploited profits to be made by discovering what is happening on other islands. Indeed the situations are not dissimilar. In our case what needs to be discovered is not only whether Q is positive or negative but also the whereabouts of its components, which are not predicted by the model.

See Also

- Business Cycles
- Loanable Funds
- Say's Law
- Temporary Equilibrium
- Trade Cycle

Bibliography

Clower, R.W. 1965. The Keynesian counterrevolution: A theoretical appraisal. In *The theory of interest rates*, ed. F.H. Hahn and F.P. Brechling. London: Macmillan.

Domar, E.D. 1946. Capital expansion, rate of growth, and employment. *Econometrica* 14: 137–147. Reprinted in *Readings in the modern theory of economic growth*, ed. J.E. Stiglitz and H. Uzawa. Cambridge, MA: MIT Press, 1969.

Goodwin, R.M. 1951. The non-linear accelerator and the persistence of business cycles. *Econometrica* 19: 1–17.

Goodwin, R.M. 1967. A growth cycle. In *Socialism, capitalism and economic growth*, ed. C.H. Feinstein. Cambridge: Cambridge University Press.

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

Harrod, R.F. 1939. An essay in dynamic theory. *Economic Journal* 49: 14–33. (Errata, June, 377.) Reprinted in *Readings in the modern theory of economic growth*, ed. J.E. Stiglitz and H. Uzawa. Cambridge, MA: MIT Press, 1969.

Harrod, R.F. 1948. *Towards a dynamic economics*. London: Macmillan.

Harrod, R.F. 1952. *Economic essays*. London: Macmillan.

Hawtrey, R.G. 1928. *Trade and credit*. London: Longmans, Green. Reprinted in American Economic Association, *Readings in business cycle theory.* Philadelphia: The Blakiston Company, 1944.

Hayek, F.A. 1935. *Prices and production*, 2nd ed. London: George Routledge & Sons.

Hicks, J.R. 1937. Mr Keynes and the 'Classics'; a suggested interpretation. *Econometrica* 5: 147–159. Republished in American Economic Association, *Readings in the theory of income distribution.* Philadelphia: The Blakiston Company, 1946

Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.

Hicks, J.R. 1950. *A contribution to the theory of the trade cycle*. Oxford: Clarendon Press.

Hicks, J.R. 1965. *Capital and growth*. Oxford: Clarendon Press.

Kaldor, N. 1940. A model of the trade cycle. *Economic Journal* 50: 78–92. Republished in N. Kaldor, *Essays on economic stability and growth.* London: Gerald Duckworth, 1960.

Kalecki, M. 1939. *Essays in the theory of economic fluctuations*. London: George Allen & Unwin.

Keynes, J.M. 1930. *A treatise on money*, vol. I. London: Macmillan.

Keynes, J.M. 1936. *The general theory of employment, interest and money.* London: Macmillan.

Keynes, J.M. 1973a. *The general theory and after. Part I. Preparation. The collected writings of John Maynard Keynes*, vol. XIII, ed. D. Moggridge. London: Macmillan.

Keynes, J.M. 1973b. *The general theory and after. Part II. Defence and development. The collected writings of John Maynard Keynes*, vol. XIV, ed. D. Moggridge. London: Macmillan.

Lucas, R.E. 1975. An equilibrium model of the business cycle. *Journal of Political Economy* 83(6): 1113–1144.

Matthews, R.C.O. 1955. The saving function and the problem of trend and cycle. *Review of Economic Studies* 22: 75–95.

Phillips, A.W. 1961. A simple model of employment, money and prices in a growing economy. *Economica* 28: 360–370.

Robertson, D.H. 1940. Mr Keynes and the rate of interest. In *Essays in monetary theory,* ed. D. H. Robertson. London: Staples Press. Republished in American Economic Association, *Readings in the theory of income distribution.* Philadelphia: The Blakiston Company, 1946.

Rose, H. 1967. On the non-linear theory of the employment cycle. *Review of Economic Studies* 34: 153–173.

Rose, H. 1969. Real and monetary factors in the business cycle. *Journal of Money, Credit, and Banking* 1(2): 138–153.

Rose, H. 1985. A policy rule for 'Say's Law' in a theory of temporary equilibrium. *Journal of Macroeconomics* 7(1): 1–17.

Samuelson, P.A. 1939. Interactions between the multiplier analysis and the principle of acceleration. In American Economic Association, *Review in business cycle theory.* Philadelphia: The Blakiston Company, 1944.

Solow, R.M. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics* 70: 65–94. Republished in *Readings in the modern theory of economic growth*, ed. J.E. Stiglitz and H. Uzawa. Cambridge, MA: MIT Press, 1969.

Aggregate Demand Theory

H. Sonnenschein

Keywords

Aggregate demand theory; Consumer demand functions; Demand theory; Individual demand functions; Market demand functions

JEL Classifications

E1

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

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

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

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

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

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

Let $\Delta^{n-1} = \{(x_1, x_2, \ldots, x_n) | x_i \geq 0$ for all i and $\sum x_i = 1\}$ Given prices p and income I, the distribution of income among consumers is defined by a mapping $\delta : R^1_{++} \times R_+ \to \Delta^{n-1}$.

Thus $\delta^i(p, I)I$ is the ith individual's income when prices are p and income is I. A candidate demand function F is a market demand function relative to the distribution of income mapping δ if there exists n consumer demand functions $f^1,\ldots, f^n$ such that $F(p, I) = \sum_i f^i[p, \delta^i(p, I)I]$ holds for all $(p,I) \in R^1_{++} \times R_+$. if $\left(f^1, \ldots f^n\right)$ are individual demand functions and if for all $\delta, \overline{\delta} \in \Delta^{n-1}$, $\sum_i f^i i \left(p, \delta^i I\right) = \sum_i \overline{f}^i i\left(p, \delta^{-i}\ I\right)$ then market demand is independent of the distribution of income.

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

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

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

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

Example 2 Let two consumers have homothetic preferences on R^2_+ represented by $U_1(x, y) = x$ and $U_2(x, y) = y$. Then market demand depends completely on the distribution of income.

If the income share of each consumer is fixed [that is, $\delta(p, I)$ is a constant vector $(\delta^1,\ldots,\delta^n)$ for all (p, I)], then homotheticity of each individual preference relation is sufficient for market demand to be utility generated. This result is due to Eisenberg (1961).

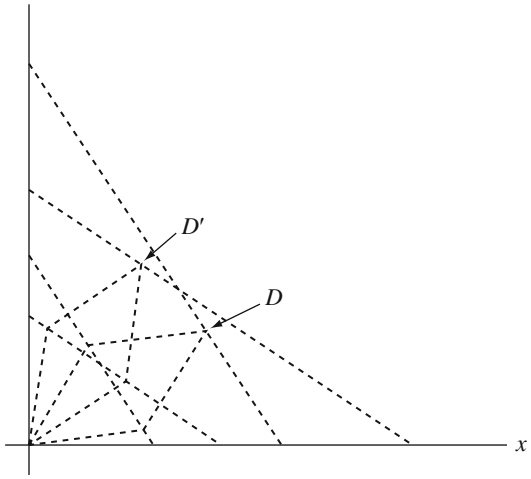
Theorem 2 (Eisenberg). If the preferences of each agent can be represented by a homogeneous of degree one utility function U^i on R^i_+, and if income shares are fixed at $(\delta^1, \ldots, \delta^n) \in \Delta^{n-1}$, then market demand is generated by the homogeneous of degree one utility function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

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

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

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

Theorems 1 and 2 referred to situations in which the distribution of income was determined exogeneously. In a much referenced paper, Samuelson (1956) presented a theorem in which the distribution of income is determined as a solution to a maximization problem. Specifically, it is assumed that for every price–income combination, the



Aggregate Demand Theory, Fig. 1

government distributes income so as to maximize a Bergsonian social welfare function: let δ denote the distribution of income function determined by this process. Samuelson's theorem asserts that under these conditions, market demand relative to δ is utility generated. Proofs of the result may also be found in Chipman and Moore (1979) and Dow and Sonnenschein (1983).

Theorem 3 Suppose that f^i is generated by U^i for $i = 1,\ldots,n$. If there exists a Bergsonian social welfare function $W(U^1,\ldots,U^n)$ that is increasing in all its arguments and such that for all $(p,I) \in R^l_{++} + R_+$,

$$\delta(p,I) \times \in \underset{(d^1,\ldots,d^n)\in\Delta^{n-1}}{\operatorname{argmax}} W\{U^1[f^1(p,d^1I)],\ldots,U^n[f^n(p,d^n,I)]\},$$

then aggregate demand $\sum_i f^i[p, \delta^i(p, I)I]$ is generated by the utility function

$$U(x) = \max W\left[U^1\left(x^1\right),\ldots,U^n(x^n)\right] \quad \text{s.t.} \quad \sum_i x^j = x.$$

3. Theorems 1–3 identify sets of assumptions under which market demand functions belong to the same class as consumer demand functions. Theorem 4 indicates that in the absence of these assumptions, none of the classical restrictions holds for market demand functions. In particular any values of demand and its derivatives that are consistent with Homogeneity and Budget balance are possible.

Theorem 4 (Sonnenschein). Let F be an arbitrary C^1 candidate demand function for l commodities and let $n \geq l$. Then, for any $(p,I) \in R^l_{++} \times R_+$ there exists a market demand function generated by n consumers with demand functions $f^1,\ldots,f^n$ such that

$$F(p,I) = \sum_{i=1}^{n} f^j\left(p, \frac{I}{n}\right)$$

and

$$\frac{\partial F_k}{\partial P_j}(p,I) = \sum \frac{\partial f^i_k}{\partial p_j}\left(p, \frac{I}{n}\right), \quad \text{for each } k,j.$$

More general results of this nature exist for market excess demand functions; see Sonnenschein (1973a), Debreu (1974) and Shafer and Sonnenschein (1982, section 4).

4. In this section an example of an economy with a continuum of infinitesimally sized agents is presented in which market demand is continuous despite the fact that individual demand functions are discontinuous: market demand is better behaved than individual demand. The point that is made here is quite general and is of importance in establishing the existence of competitive equilibrium without need for the assumption that preferences are convex; see Debreu (1982, section 4).

Example 4 There are two commodities x and y and the preferences of a consumer of type a are represented by the utility function $U(x, y, a) = x^2 + a^2 \cdot y^2$ The income of each consumer is fixed at unity and the consumption set of each consumer if R^2_+.

The price of commodity y in terms of the numeraire commodity x is denoted by p. The distribution of agent types is specified by defining the following density function g, over the domain of a:

$$g(a) = \begin{cases} 2 & \text{if } a \in \left[\frac{1}{4}, \frac{3}{4}\right] \\ 0 & \text{otherwise} \end{cases}.$$

Strict convexity of preferences is violated for each a, and consequently, the demand function of each consumer type is not single valued. The demand function for y as a function of p is given by

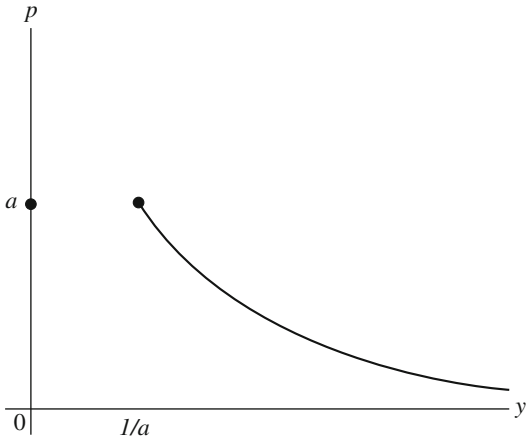
$$f^a(p) = \begin{cases} \frac{1}{p} & \text{if } p < a \\ 0 & \text{if } p > a \\ \left[\frac{1}{a}, 0\right] & \text{if } p = a. \end{cases}$$

The graph of f^a is drawn in Fig. 2.

The multi-valued function f^a is not well-behaved in the sense that it jumps at a.

Let $F(p)$ denote market demand at price p. By definition

$$\begin{aligned} F(p) &= 2\int_{a=1/4}^{a=3/4} f^a(p)\mathrm{d}a \\ &= 2\int_{a=1/4}^{a=p} (0)\mathrm{d}a + 2\int_{a=p}^{a=3/4} \frac{1}{p}\ \mathrm{d}a \\ &= \frac{3}{2p} - 2. \end{aligned}$$



Aggregate Demand Theory, Fig. 2

Thus, market demand is single-valued and differentiable in the entire domain of p, despite the fact that these properties do not hold for any given a. One way to understand the result is to observe that for each p, the relative mass of consumers whose demand is discontinuous at p is zero. This observation also illustrates the importance of the assumption that each agent is a 'small' part of the market and that preferences are dispersed. The result would not hold if the density function was assumed to be

$$h(a) = \begin{cases} 1 & \text{if } a \in \left[\frac{1}{4}, \frac{3}{4}\right] \\ \frac{1}{2} & \text{if } a = 1 \\ 0 & \text{otherwise} \end{cases}$$

A final result, which illustrates a theorem due to Hildenbrand (1983), gives conditions under which market demand is necessarily downward sloping. Again, the point is that with the continuum of agents market demand may be better behaved than individual demand.

Theorem 5 Consider an economy in which all individuals have identical preferences but differ in their incomes. In particular, assume that income is uniformly distributed over the interval [0, 1] and let $f(p, I)$ denote the identical demands of the individuals with income I who face prices p. Under the above conditions, the mean demand for each commodity has a nonpositive slope.

A sketch of a proof of the theorem follows: It is well known from consumer demand theory that the sign of the term $\partial f_k(p, I)/\partial p_k$ can be either positive or negative. Since individual substitution effects are nonpositive, to prove the result it is sufficient to demonstrate that the mean income effect is nonpositive.

The income effect as a result of a change in the price of commodity k on the demand for k, for an individual with income I, is given by

$$-f_k(p,I)\frac{\partial}{\partial I} f_k(p,I).$$

Therefore, the mean income effect is given by

$$-\int_0^1 f_k(p,I)\frac{\partial}{\partial I}f_k(p,I)\mathrm{d}I = -\frac{1}{2}\int_0^1 \frac{\partial}{\partial I}\left[f_k{}^2(p,I)\right]\mathrm{d}I$$
$$= -\frac{1}{2}\left[f_k{}^2(p,I) - f_k{}^2(p,0)\right] = -\frac{1}{2}f_k{}^2(p,1) \le 0,$$

which establishes the result.

See Also

- ▶ Aggregation (Theory)
- ▶ Demand Theory
- ▶ Integrability of Demand
- ▶ Law of Demand

Bibliography

Antonelli, G.B. 1886. *Sulla Teoria Matematica della Economia Politica*. Pisa: Nella Tipognafia del Folchetto. Trans. J.S. Chipman and A.P. Kirman in *Preferences, utility and demand*, ed. J.S. Chipman et al., New York: Harcourt Brace Jovanovich, 1971.

Arrow, K.J. and Intriligator, M.D., eds. 1981–1986. *Handbook of mathematical economics*, Vols 1, 2 and 3. Amsterdam: North-Holland.

Chipman, J.S., and J. Moore. 1979. On social welfare functions and the aggregation of preferences. *Journal of Economic Theory* 21: 111–139.

Chipman, J.S., L. Hurwicz, M.K. Richter, and H.F. Sonnenschein, eds. 1971. *Preferences, utility and demand*. New York: Harcourt Brace Jovanovich.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1: 15–23.

Debreu, G. 1982. Existence of competitive equilibrium. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

Dow, J., and H. Sonnenschein. 1983. Samuelson and Chipman–Moore on utility generated community demand. In *Prices, competition and equilibrium*, ed. M. Peston and R. Quandt. Oxford: Philip Allan Publishers.

Eisenberg, B. 1961. Aggregation of utility functions. *Management Science* 7: 337–350.

Gorman, W.M. 1953. Community preference fields. *Econometrica* 21: 63–80.

Hicks, J.R. 1957. *A revision of demand theory.* Oxford: Clarendon Press.

Hildenbrand, W. 1983. On the 'law of demand'. *Econometrica* 51: 997–1020.

Hurwicz, L., and H. Uzawa. 1971. On the integrability of demand functions. In *Preferences, utility and demand*, ed. J.S. Chipman et al. New York: Harcourt Brace Jovanovich.

Nataf, A. 1953. Sur des qsts d'agrégation en économétrie. *Publications de l'Institute de Statistique de l'Université de Paris* 2: 5–61.

Samuelson, P. 1956. Social indifference curves. *Quarterly Journal of Economics* 70: 1–22.

Shafer, W., and H. Sonnenschein. 1982. Market demand and excess demand functions. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator. Amsterdam: North-Holland.

Sonnenschein, H. 1973a. Do Walras' identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6: 345–354.

Sonnenschein, H. 1973b. The utility hypothesis and market demand theory. *Western Economic Journal* 11: 404–410.

A

Aggregate Supply Function

Paul Davidson

John Maynard Keynes wrote *The General Theory* (1936) in order to show that Say's Law, where (aggregate) supply created its own (aggregate) demand, was not applicable to a monetary, production economy. In a Say's Law world, the aggregate demand function would be coincident with the aggregate supply function so that 'effective demand, instead of having a unique equilibrium value, is an infinite range of values all equally admissible; and the amount of employment is indeterminate except in so far as the marginal disutility of labour sets an upper limit' (Keynes 1936, p. 26). In other words, Say's Law assumes there is no barrier to the economy obtaining, in the long run, a full employment output level.

Keynes claimed that Say's Law 'is not the true law relating the aggregate demand and supply functions' (1936, p. 26) and hence the 'true' relationship between the aggregate demand and the aggregate supply functions 'remains to be written and without which all discussions concerning the volume of aggregate employment are futile' (1936, p. 26). As Keynes pointed out in a letter to D. H. Robertson (Keynes 1973), however, his aggregate supply function was 'simply the

age-old supply function'. Keynes's revolutionary analysis stemmed from his belief that in a monetary economy, the aggregate demand function differed from, and was *not* coincident with, the aggregate supply function.

Keynes argued that the aggregate supply function could be readily derived from ordinary Marshallian micro-supply functions (1936, pp. 44–5) and that, therefore, the properties of the aggregate supply function 'involved few considerations which are not already familiar' (1936, p. 89). Keynes believed that 'it was the part played by the aggregate demand function which has been overlooked' (1936, p. 89). Hence, though Keynes briefly described the aggregate supply function (1936, pp. 25, 44–5) and its inverse, the employment function (1936, pp. 89, 280–1), the bulk of *The General Theory* was devoted to developing the characteristics of aggregate demand while the aggregate supply function was treated perfunctorily.

Consequently, the 'Keynesian Revolution' analytical structure (which Samuelson dubbed 'neoclassical synthesis Keynesianism') which was developed by Hicks (1937), Modigliani (1944), and Klein (1947) emphasized the novelty of the aggregate demand-side of Keynes's economic system. In losing sight of Keynes's well-known 'age-old' aggregate supply function, the Keynesian Revolution went off half-cocked and lost its foundation in Marshallian microeconomics.

In the 1954–7 period, there was a flurry of activity attempting to rediscover the basis of Keynes's aggregate supply function. This discussion culminated in Weintraub's 1957 article which Clower, in personal correspondence (dated 1 November 1957), characterized as 'a beautifully clear statement of what Keynes "should have meant" if we suppose that he was a rational being'.

The aggregate supply function as stated by Keynes and explicitly developed by Weintraub (1957), Davidson (1962), and Davidson and Smolensky (1964) relates the aggregate number of workers (N) that profit-maximizing entrepreneurs would want to hire for each possible level of expected sales proceeds (Z)-given the money wage rate, technology, the degree of competition (or monopoly), and the degree of integration of firms (cf. Keynes 1936, p. 245). For any given degree of firm integration in the aggregate, GNP is directly related to total sales proceeds. If firms are fully integrated, aggregate sales proceeds equals GNP.

Following Keynes's argument (1936, p. 41) that money values and quantities of employment are the two 'fundamental units of quantity' to be used when dealing with aggregates, the aggregate supply proceeds are normally specified either in money terms (Z) or in Keynes's wage unit terms (Z_w) which is money sales proceeds divided by the money wage rate. Hence the aggregate supply function is specified as:

$$Z = f_1(N) \tag{1}$$

or

$$Z_w = f_2(N) \tag{2}$$

For purposes of simplicity and ease of comparability with the ordinary Marshallian micro-supply function, only the form of Eq. 1 will be developed in the following discussion. Equational form (2) of the aggregate supply function can then be derived merely by dividing all money sums expressed in Eq. 1 by the existing money wage rate.

The Marshallian supply curve for a single firm (s_f) indicates the profit-maximizing output possibilities for alternative market demand conditions. The supply schedule of profit-maximizing, alternative price-quantity combinations depend on the degree of competition (or monopoly) of the firm (k) and its marginal costs (MC).

The degree of monopoly of the firm depends on the market demand condition it faces. In the most simple case, as aggregate demand changes the demand curve facing the firm shifts without altering the degree of monopoly of the firm; for example, in the perfectly competitive case, shifts in the firm's demand curve do not alter the competitive market conditions. In more complex cases the degree of monopoly may vary as aggregate demand changes and the firm's demand curve shifts, i.e. $k = f(N)$.

Thus the firm's supply schedule can be specified in terms of its degree of monopoly power as

given by a mark-up – whose magnitude depends on the price elasticity of demand facing the firm and its marginal costs:

$$s_f = f_3\left(k_f, MC_f\right) \tag{3}$$

where K_f is the firm's mark-up over its marginal costs (MC_f).

The profit-maximizing firm's mark-up is equal to Lerner's (1934) measure of the degree of monopoly power which is $[1/E_{\mathrm{df}}]$ where E_{df} is the price elasticity of demand facing the firm for any given level of effective demand. Thus, for a perfectly competitive firm, $k = 0$ for all potential production flows and only marginal costs affect the position and shape of its marginal cost curve. For conditions of less than perfect competition, $k > 0$, and hence both marginal costs and monopoly power at each potential output level affect the firm's market offerings as reflected in its supply curve offerings.

The firm's marginal cost (MC_f), assuming labour is the only variable input in the production process, equals the money wage (w) divided by marginal labour productivity (MP) where the latter is a function of employment (and the laws of returns involved in the technology of the firm). For any given 'law of returns' facing the firm, there will be a different marginal production cost structure. For example, with diminishing returns, the marginal production costs increase with increasing output; for constant returns, marginal production costs are constant, while for decreasing returns marginal costs decline with increases in output and employment. (Of course, the latter case is incompatible with perfect competition; it requires some degree of monopoly and hence some positive mark-up, $[k > 0]$ over marginal costs, so that market price covers average unit costs). If marginal user costs (MUC) are not negligible, then $MC_f = [w/MP + MUC]$.

The Marshallian industry flow-supply schedule (s) is obtained simply by the usual lateral summation of the individual firm's supply curves; it is, therefore, related to the average industry mark-up or 'average' degree of monopoly *and* the industry's marginal cost schedule, i.e.,

$$s = f_4[k, MC] \tag{4}$$

where the symbols without subscripts are the industry's equivalent to the aforementioned firm's variables. Thus given (a) the production technology, (b) the money wage, and (c) the degree of monopoly based on specified market conditions for any given potential output and employment level, a unique industry supply function can be derived.

Although output across firms in the same industry may be homogeneous and therefore can be aggregated to obtain the industry supply schedule (Eq. 4), this homogeneity of output assumption cannot be accepted as the basis for summing across industries to obtain the aggregate supply function (Keynes 1936, ch. 4). Accordingly, the Marshallian industry supply function, s, which relates prices (p) and quantities (q) must be transformed into Keynes's industry supply function which relates total industry sales proceeds in money terms (z) with total industry employment hiring (n), i.e.,

$$z = f_5(n) \tag{5}$$

Since given returns, the money-wage, and the degree of monopoly, every point on the Marshallian industry supply function, s, is associated with a unique profit-maximizing price-quantity combination whose multiple equals total expected sales proceeds (i.e., $p \times q = z$) and since every industry output level (q) can be associated with a unique industry hiring level, i.e. $q = f(n)$, then every point of Eq. 4 of the s-curve in p-q space can be transformed to a point on a z-curve in pq-n space to obtain Eq. 5 *supra*.

Hence for each industry in which the traditional Marshallian supply function can be formulated in terms of Eq. 4, a Keynes industry supply function (Eq. 5) can also be uniquely specified. All of Keynes's industry supply functions can then be aggregated together to obtain the aggregate supply function in terms of aggregate money proceeds (Z) and the aggregate quantity of employment units (N) as specified in Eq. 1, provided one reasonably assumes that corresponding to any given point of aggregate supply there is a

unique distribution of proceeds and employment between the different industries in the economy (Keynes 1936, p. 282).

See Also

- ▶ Keynes, John Maynard (1883–1946)
- ▶ Keynes's General Theory

Bibliography

Davidson, P. 1962. More on the aggregate supply function. *Economic Journal* 72: 452–457.
Davidson, P., and E. Smolensky. 1964. *Aggregate supply and demand analysis*. New York: Harper & Row.
Hicks, J.R. 1937. Mr Keynes and the classics; a suggested interpretation. *Econometrica* 5: 147–159.
Keynes, J.M. 1936. *The general theory of employment, interest and money*. New York: Harcourt, Brace.
Keynes, J.M. 1973. In *The collected writings of John Maynard Keynes*, vol. XIII, ed. D. Moggridge. London: Macmillan.
Klein, L.R. 1947. *The Keynesian revolution*. New York: Macmillan.
Lerner, A.P. 1934. The concept of monopoly and the measurement of monopoly power. *Review of Economic Studies* 1: 157–175.
Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12: 45–88.
Weintraub, S. 1957. The micro-foundations of aggregate demand and supply. *Economic Journal* 67: 455–470.

Aggregation (Econometrics)

Thomas M. Stoker

Abstract

The econometrics of aggregation is about modelling the relationship between individual (micro) behaviour and aggregate (macro) statistics, so that data from both levels can be used for estimation and inference about economic parameters. Practical models must address three types of individual heterogeneity – in income and preferences, in wealth and income risk, and in market participation. This entry discusses recent solutions to these problems in the context of demand analysis, consumption modelling and labour supply. Also discussed is work that uses aggregation structure to solve microeconometric estimation problems, and work that addresses whether macroeconomic interactions provide approximate solutions to aggregation problems.

Keywords

Aggregate demand models; Aggregation (econometrics); Aggregation factors; Approximate aggregation; Calibration; Computable stochastic growth models; Constant relative risk aversion; Demand models; Exact aggregation; Gorman, W. (Terence); Household demand models; Identification; Income-risk insurance; Individual heterogeneity; Industrial organization; Law of demand; Mills ratio; Reservation wage; Selection effects; Theil, H.; Uncorrelated transfers

JEL Classifications

C43

Aggregation refers to the connection between economic interactions at the micro and the macro levels. The micro level refers to the behaviour of individual economic agents. The macro level refers to the relationships that exist between economy-wide totals, averages or other economic aggregates. For instance, in a study of savings behaviour refers to the process that an individual or household uses to decide how much to save out of current income, whereas the aggregates are total or per-capita savings and income for a national economy or other large group. The econometrics of aggregation refers to modelling with the individual–aggregate connection in mind, creating a framework where information on individual behaviour together with co-movements of aggregates can be used to estimate a consistent econometric model.

In economic applications one encounters many types and levels of aggregation: across goods, across individuals within households, and so on. We focus on micro to macro as outlined

above, and our 'individual' will be a single individual or a household, depending on the context. We hope that this ambiguity does not cause confusion.

At a fundamental level, aggregation is about handling detail. No matter what the topic, the microeconomic level involves purposeful individuals who are dramatically different from one another in terms of their needs and opportunities. Aggregation is about how all this detail distils in relationships among economic aggregates.

Understanding economic aggregates is essential for understanding economic policy. There is just too much individual detail to conceive of tuning policies to the idiosyncrasies of many individuals.

This detail is referred to as individual heterogeneity, and it is pervasive. This is a fact of empirical evidence and has strong econometric implications. If you ignore or neglect individual heterogeneity, then you can't get an interpretable relationship between economic aggregates. Aggregates reflect a smear of individual responses and shifts in the composition of individuals in the population; without careful attention, the smear is unpredictable and uninterpretable.

Suppose that you observe an increase in aggregate savings, together with an increase in aggregate income and in interest rates. Is the savings increase primarily arising from wealthy people or from those with moderate income? Is the impact of interest rates different between the wealthy and others? Is the response different for the elderly than for the young? Has future income for most people become more risky?

How could we answer these questions? The change in aggregate savings is a mixture of the responses of all the individuals in the population. Can we disentangle it to understand the change at a lower level of detail, like rich versus poor, or young versus old? Can we count on the mixture of responses underlying aggregate savings to be stable? These are questions addressed by aggregation.

Recent progress on aggregation and econometrics has centred on explicit models of individual heterogeneity. It is useful to think of heterogeneity as arising from three broad categories of differences. First, individuals differ in tastes and incomes. Second, individuals differ in the extent to which they participate in markets. Third, individuals differ in the situations of wealth and income risk that they encounter depending on the market environment that exists. Our discussion of recent solutions is organized around these three categories of heterogeneity. For deeper study and detailed citations, see the surveys by Blundell and Stoker (2005), Stoker (1993) and Browning et al. (1999).

The classical aggregation problem provides a useful backdrop for understanding current solutions. We now review its basic features, as originally established by Gorman (1953) and Theil (1954). Suppose we are studying the consumption of some product by households in a large population over a given time period t. Suppose that the quantity purchased q_{it} is determined by household resources m_{it}, or 'income' for short, as in the formula:

$$q_{it} = \alpha_i + \beta_i m_{it}$$

Here α_i represents a base level consumption, and β_i represents household i's marginal propensity to spend on the product.

For aggregation, we are interested in what, if any, relationship there is between average quantity and average income:

$$\overline{q}_t = \frac{1}{n_t}\sum_{i=1}^{n_t} q_{it} \text{ and } \overline{m}_t = \frac{1}{n_t}\sum_{i=1}^{n_t} m_{it}$$

where all households have been listed as $i = 1,\ldots, n_t$. Let's focus on one version of this issue, namely, what happens if some new income becomes available to households, either through economic growth or a policy. How will the change in average quantity purchased $\Delta\overline{q}$ be related to the change in average income $\Delta\overline{m}$?

Suppose that household i gets Δm_i in new income. Their change in quantity purchased is the difference between purchases at income $m_{it} + \Delta m_i$ and at income m_{it}, or

$$\Delta q_i = \beta_i \cdot \Delta m_i$$

Now, the average quantity change is $\Delta\overline{q} = \sum_i \Delta q_i / n_t$, so that

$$\Delta\overline{q} = \frac{1}{n_t}\sum_{i=1}^{n_t}\beta_i \cdot \Delta m_i \tag{1}$$

In general, it seems we need to know a lot about who gets the added income – which i's get large values of Δm_i and which i's get small values of Δm_i. With a transfer policy, any group of households could be targeted for the new income, and their specific set of values of β_i would determine $\Delta\overline{q}$. A full schedule of how much new income goes to each household i as well as how they spend it (that is, Δm_i and β_i), seems like a lot of detail to keep track of, especially if the population is large.

Can we ever get by knowing just the change in average income $\Delta\overline{m} = \sum_i \Delta m_i / n_t$?

There are two situations where we can, where a full schedule is not needed:

1. Each household spends in exactly the same way, namely, $\beta_I = \beta$ for all i, so that who gets the new income doesn't affect $\Delta\overline{q}$.
2. The distribution of income transfers is restricted in a convenient way.

Situation 1 is (common) micro linearity, which is termed *exact aggregation*. Another way to understand the structure is to write (1) in the covariance formulation:

$$\Delta\overline{q} = \overline{\beta}\cdot\Delta\overline{m} + \frac{1}{n_t}\sum_{i=1}^{n_t}\left(\beta_i - \overline{\beta}\right)\cdot\left(\Delta m_i - \Delta\overline{m}\right) \tag{2}$$

where we denote the average spending propensity as $\overline{\beta} = \sum_i \beta_i / n_t$. With exact aggregation there is no variation in β_i, so that $\beta_i = \beta = \overline{\beta}$ and the latter term always vanishes. That is, it doesn't matter who gets the added income because everyone spends the same way. When there is variation in β_i, matters are more complicated unless it can be assured that the new income were always given to households in a way that is uncorrelated with the propensities β_i. 'Uncorrelated transfers' provide an example of a Situation 2, but that is a distribution restriction that is hard to verify with empirical data.

Under uncorrelated transfers, we can also interpret the relationship between $\Delta\overline{q}$ and $\Delta\overline{m}$, that is, the macro propensity is the average propensity $\overline{\beta}$. There are other distributional restrictions that give a constant macro propensity, but a different one from the parameter produced by uncorrelatedness. For instance, suppose that transfers of new income always involved fixed shares of the total amount. That is, household i gets

$$\Delta m_i = s_i \Delta\overline{m} \tag{3}$$

In this case, average purchases are

$$\Delta\overline{q} = \frac{1}{n_t}\sum_{i=1}^{n_t}\beta_i \cdot (s_i\Delta\overline{m}) = \tilde{\beta}_{wtd}\cdot\Delta\overline{m} \tag{4}$$

where $\tilde{\beta}_{wtd}$ is the weighted average $\tilde{\beta}_{wtd} \equiv \sum_i \beta_i$ s_i / n_t. This is a simple aggregate relationship, but the coefficient $\tilde{\beta}_{wtd}$ applies only for the distributional scheme (3); it matters who gets what share of the added income. Aside from being a weighted average of $\{\beta_i\}$, there is no reason for $\tilde{\beta}_{wtd}$ to be easily interpretable – for instance, if households with low β_i's have high s_i's, then $\tilde{\beta}_{wtd}$ will be low. If your aim was to estimate the average propensity β, there is no reason to believe that the bias $\tilde{\beta}_{wtd} - \overline{\beta}$ will be small.

Empirical models that take aggregation into account apply structure to individual responses and to allowable distributional shifts. Large populations are modelled, so that compositional changes are represented via probability distributions, and expectations are used instead of averages (for example, mean quantity $E_t(q)$ is modelled instead of the sample average $\overline{q}_t$). Individual heterogeneity is the catch-all term for individual differences, and they must be characterized. Distribution restrictions must be applied where heterogeneity is important. For instance, in our example structure on the distribution of new income is required for dealing with the heterogeneity in β_i, but not for the heterogeneity in α_i.

Progress in empirical modelling has come about because of the enhanced availability of micro data over time. The forms of behavioural models in different research areas have been tightly characterized, which is necessary for understanding how to account for aggregation. That is, when individual heterogeneity is

characterized empirically, the way is clear to understanding what distributional influences are relevant and must be taken into account. We discuss recent examples of this below.

Some Solutions to Aggregation Problems

Demand Models and Exact Aggregation

It is well known that demand patterns of individual households vary substantially with whether households are rich or poor, and vary with many observable demographic characteristics, such as household (family) size, age of head and ages of children, and so on. As surveyed in Blundell (1988), traditional household demand models relate household commodity expenditures to price levels, total household budget (income) and observable household characteristics. Aggregate demand models relate (economy-wide) aggregate commodity expenditures to price levels and the distribution of income and characteristics in the population. Demand models illustrate exact aggregation, a practical approach for accommodating heterogeneity at the micro and macro levels. These models assume that demand parameter values are the same for all individuals, but explicitly account for observed differences in tastes and income.

For instance, suppose we are studying the demand for food and we are concerned with the difference in demands for households of small size versus large size. We model food purchases for household i as part of static allocation of the budget m_{it} to $j = 1,...,J$ expenditure categories, where food is given by $j = 1$, and price levels at time t are given by $P_t = (p_{1t},...,p_{Jt})$. Small families are indicated by $z_{it} = 0$ and large families by $z_{it} = 1$.

Expenditure patterns are typically best fit in budget share form. For instance, a translog model of the food share takes the form

$$w_{1it} = \frac{p_1 q_{1it}}{m_{it}} = \frac{1}{D(p_t)}\left[\alpha_1 + \sum_{i=1}^{J}\beta_{1j}\ln p_{jt} + \beta_m \ln m_{it} + \beta_z z_{it}\right] \quad (5)$$

where $D(p_t) = 1 + \sum_{i=1}^{J} \beta_j \ln p_{jt}$. The parameters ($\alpha_1$ and all β's) are the same across households, and the price levels (p_{jt}'s) are the same for all households but vary with t. Individual heterogeneity is represented by the budget m_{it} and the family size indicator z_{it}. We have omitted an additive disturbance for simplicity, which would represent another source of heterogeneity. The important thing for aggregation is that model (5) is intrinsically linear in the individual heterogeneity. That is, we can write

$$w_{1it} = b_1(p_t) + b_m(p_t)\cdot \ln m_{it} + b_z(p_t)\cdot z_{it} \quad (6)$$

The aggregate share of food in the population is the mean of food expenditures divided by mean budget, or

$$W_{1t} = \frac{E_t(m_{it}w_{1it})}{E_t(m_{it})} = b_1(p_t) + b_m(p_t)\cdot\frac{E_t\,(m_{it}\ln m_{it})}{E_t(m_{it})} + b_z(p_t)\cdot\frac{E_t\,(m_{it}z_{it})}{E_t(m_{it})} \quad (7)$$

The aggregate share depends on prices, the parameters (α_1 and all β's) and two statistics of the joint distribution of m_{it} and z_{it}. The first,

$$S_{mt} = \frac{E_t\,(m_{it}\ln m_{it})}{E_t(m_{it})} \quad (8)$$

is an entropy term that captures the size distribution of budgets, and the second

$$S_{zt} = \frac{E_t\,(m_{it}z_{it})}{E_t(m_{it})} \quad (9)$$

is the percentage of total expenditure accounted for by households with $z_{it} = 1$, that is, large families.

The expressions (6) and (7) illustrate *exact aggregation* models. Heterogeneity in tastes and budgets (incomes) are represented in an intrinsically linear way. For aggregate demand, all one needs to know about the joint distribution of budgets m_{ti} and household types z_{it} is a few statistics; here S_{mt} and S_{zt}.

The obvious similarity between the individual model (6) and the aggregate model (7) raises a further question. How much bias is introduced by just fitting the individual model with aggregate data, that is, putting $E_t(m_{it})$ and $E_t(z_{it})$ in place of m_{it} and z_{it}, respectively? This can be judged by the use of *aggregation factors*. Define the factors π_{mt} and π_{zt} as

$$\pi_{mt} = \frac{S_{mt}}{\ln E(m_{it})} \text{ and } \pi_{zt} = \frac{S_{zt}}{E_t(z_{it})}$$

so that the aggregate share is

$$\begin{aligned} W_{1t} &= \frac{E_t(m_{it} w_{1it})}{E_t(m_{it})} \\ &= b_1(p_t) + b_m(p_t) \cdot \pi_{mt} \cdot \ln E_t(m_{it}) \\ &\quad + b_z(p_t) \cdot \pi_{zt} \cdot E_t(z_{it}) \end{aligned}$$

One can learn about the nature of aggregation bias by studying the factors π_{mt} and π_{zt}. If they are both roughly equal to 1 over time, then no bias would be introduced by fitting the individual model with aggregate data. If they are roughly constant but not equal to 1, then constant biases are introduced. If the factors are time varying, more complicated bias would result. In this way, with exact aggregation models, aggregation factors can depict the extent of aggregation bias.

The current state of the art in demand analysis uses models in exact aggregation form. The income (budget) structure of shares is adequately represented as quadratic in $\ln m_{it}$, as long as many demographic differences are included in the analysis. This means that aggregate demand depends explicitly on many statistics of the income-demographic distribution, and it is possible to gauge the nature and sources of aggregation bias using factors as we have outlined. See Banks et al. (1997) for an example of demand modelling of British expenditure data, including the computation of various aggregation factors.

Exact aggregation modelling arises naturally in situations where linear models have been found to provide adequate explanations of empirical data patterns. This is not always the case, as many applications require models that are intrinsically nonlinear. We now discuss an example of this kind where economic decisions are discrete.

Market Participation and Wages

Market participation is often a discrete decision. Labourers decide whether to work or not, firms decide whether to enter a market or exit a market. There is no 'partial' participation in many circumstances, and changes are along the extensive margin. This raises a number of interesting issues for aggregation.

We discuss these issues using a simple model of labour participation and wages. We consider two basic questions. First, how is the fraction of working (participating) individuals affected by the distribution of factors that determine whether each individual chooses to work? Second, what is the structure of average wages, given that wages are observed only for individuals who choose to work? The latter question is of interest for interpreting wage movements: if average wages go up, is that because (*a*) most individual wages went up or (*b*) low-wage individuals become unemployed, or leave work? These two reasons give rise to quite different views of the change in economic welfare associated with an increase in average wages.

The standard empirical model for individual wages expresses log wage as a linear function of time effects, schooling and demographic (cohort) effects. Here we begin with

$$\ln w_{it} = r(t) + \beta \cdot S_{it} + \varepsilon_{it} \tag{10}$$

where $r(t)$ represents a linear trend or other time effects, S_{it} is the level of training or schooling attained by individual i at time t, and ε_{it} are all other idiosyncratic factors. This setting is consistent with a simple skill price model, where $w_{ti} = R_t H_{it}$ with skill price $R_t = e^{r(t)}$ and skill (human capital) level $H_{it} = e^{\beta S_{it} + \varepsilon_{it}}$. We take Eq. (10) to apply to all individuals, with the wage representing the available or offered wage, and β the return to schooling. However, we observe that wage only for individuals who choose to work.

We assume that individuals decide whether to work by first forming a reservation wage

$$\ln w_{it}^* = s^*(t) + \alpha \ln B_{it} + \beta^* \cdot S_{it} + \zeta_{it}$$

where $s(t)$ represents time effects, B_{it} is the income or benefits available when individual i is out of work at time t, S_{it} is schooling as before, and ζ_{it} are all other individual factors. Individual i will work at time t if their offered wage is as big as their reservation wage, or $w_{it} \geq w_{it}^*$. We denote this by the participation indicator I_{it}, where $I_{it} = 1$ if i works and $I_{it} = 0$ if i doesn't work. This model of participation can be summarized as

$$\begin{aligned} I_{it} &= 1\left[w_{it} \geq w_{it}^*\right] = 1\left[\ln w_{it} - \ln w_{it}^* \geq 0\right] \\ &= 1\left[s(t) - \alpha \ln B_{it} + \gamma \cdot S_{it} + \nu_{it} \geq 0\right] \end{aligned} \quad (11)$$

where $s(t) \equiv r(t) - s^*(t)$, $\gamma \equiv \beta - \beta^*$ and $\nu_{It} \equiv \varepsilon_{it} - \zeta_{it}$.

If the idiosyncratic terms ε_{it}, ν_{it} are stochastic errors with zero means (conditional on B_{it},S_{it}) and constant variances, then (10) and (11) is a standard selection model. That is, if we observe a sample of wages from working individuals, they will follow (10) subject to the proviso that $I_{it} = 1$. This can be accommodated in estimation by assuming that ε_{it}, ν_{it} have a joint normal distribution. That implies that the log wage regression of the form (10) can be corrected by adding a standard selection term as

$$\begin{aligned} \ln w_{it} &= r(t) + \beta \cdot S_{it} \\ &+ \frac{\sigma_{\varepsilon v}}{\sigma_v} \lambda\left[\frac{s(t) - \alpha \ln B_{it} + \gamma S_{it}}{\sigma_v}\right] + \eta_t. \end{aligned} \quad (12)$$

Here, σ_v is the standard deviation of v and $\sigma_{\varepsilon v}$ is the covariance between ε and v. $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the 'Mills ratio', where $\phi$and Φ are the standard normal p.d.f. and c.d.f respectively. This equation is properly specified for a sample of working individuals – that is, we have $E(\eta_t|S_{it}, B_{it}, I_{it} = 1) = 0$. For a given levels of benefits and schooling, Eq. (11) gives the probability of participating in work as

$$E_t\left[I_{it}|B_{it}, S_{it}\right] = \Phi\left[\frac{s(t) - \alpha \ln B_{it} + \gamma \cdot S_{it}}{\sigma_v}\right] \quad (13)$$

where $\Phi[\cdot]$ is the normal c.d.f.

For studying average wages, the working population is all individuals with $I_{it} = 1$. The fraction of workers participating is therefore the (unconditional) probability that $\alpha \ln B_{it} - \gamma \cdot S_{it} - \nu_{it} \leq s(t)$. This probability is the expectation of I_{it} in (11), an intrinsically nonlinear function in observed heterogeneity B_{it} and S_{it} and unobserved heterogeneity ν_{it}, so we need some explicit distribution assumptions. In particular, assume that the participation index $\alpha \ln B_{it} - \gamma \cdot S_{it} - \nu_{it}$ is normally distributed with mean $\mu_t = \alpha E_t(\ln B_{it}) - \gamma E_t(S_{it})$ and variance

$$\begin{aligned} \sigma_t^2 &= \alpha^2 Var_t(\ln B_{it}) + \beta^2 Var_t(S_{it}) - 2a\beta \\ &\cdot Cov_t(\ln B_{it}, S_{it}) + \sigma_v^2. \end{aligned} \quad (14)$$

Now we can derive the labour participation rate (or one minus the unemployment rate) as

$$E_t[I_{it}] = \Phi\left[\frac{s(t) - \alpha E_t(\ln B_{it}) + \gamma E_t(S_{it})}{\sigma_t}\right] \quad (15)$$

where again $\Phi[\cdot]$ is the normal c.d.f. This formula relates the participation rate to average out-of-work benefits $E_t(\ln B_{it})$ and average training $E_t(S_{it})$, as well as their variances and covariances through σ_t. The specific relation depends on the distributional assumption adopted; (15) relies on normality of the participation index in the population.

For wages, a similar analysis applies. Log wages are a linear function (10) applicable to the full population. However, for participating individuals, the intrinsically nonlinear selection term is introduced, so that we need explicit distributional assumptions. Now suppose that log wage $\ln w_{it}$ and the participation index $\alpha \ln B_{it} - \gamma \cdot S_{it} - \nu_{it}$ are joint normally distribution. It is not hard to derive the expression for average log wages of working individuals

$$\begin{aligned} E_t[\ln w_{it}|I_{it} = 1] = \; & r(t) + \beta \cdot E_t(S_{it}|I_{it} = 1) \\ & \frac{\sigma_{\varepsilon v}}{\sigma_t} \lambda\left[\frac{s(t) - \alpha E_t(\ln B_{it}) + \gamma E_t(S_{it})}{\sigma_t}\right]. \end{aligned} \quad (16)$$

This is an interesting expression, which relates average log wage to average training of the

workers as well as to the factors that determine participation.

However, we are not interested in average log wages, but rather average wages $E_t(w_{it})$. The normality structure we have assumed is enough to derive a formulation of average wages, although it is a little complex to reproduce in full here. In brief, Blundell et al. (2003) show that the average wages of working individuals $E[w_{it}|I_{it}=1]$ can be written as

$$\ln E[w_{it}|I_{it}=1] = r(t) + \beta \cdot E_t(S_{it}) + \Omega_t)\Psi_t \quad (17)$$

where Ω_t, Ψ_t are correction terms that arise as follows. Ω_t corrects for the difference between the log of an average and the average of a log, as

$$\Omega_t = \ln E_t(w_{it}) - E_t(\ln w_{it}) + \Omega_t .$$

Ψ_t corrects for participation, as

$$\Psi_t \equiv \ln E[w_{it}|I_{it}=1] - \ln E_t(w_{it}).$$

Recall our original question, about whether an increase in average wages is due to an increase in individual wages or to increased unemployment of low-wage workers. That is captured in (17). That is, Ψ_t gives the participation effect, and the other terms capture changes in average wage $E_t(w_{it})$ when all are participating. As such, this analysis provides a vehicle for separating overall wage growth from compositional effects due to participation.

Blundell et al. (2003) analyse British employment using a framework similar to this, but also allowing for heterogeneity in hours worked. Using out-of-work benefits as an instrument for participation, they find that over 40 per cent of observed aggregate wage growth from 1978 to 1996 arises from selection and other compositional effects.

We have now discussed aggregation and heterogeneity with regard to tastes and incomes, and market participation. We now turn to heterogeneity with regard to risks and market environments.

Consumption and Risk Environments

Consumption and savings decisions are clearly affected by preference heterogeneity, as we discussed earlier. The present spending needs of a large family clearly differ from those of a small family or a single individual, the needs of teenage children differ from those of preschoolers, the needs of young adults differ from those of retirees, and on and on. These aspects are very important, and need to be addressed as they were in demand models above. Browning and Lusardi (1996) survey the extensive evidence on heterogeneity in consumption, and Attanasio (1999) is an excellent comprehensive survey of work on consumption.

We use consumption and savings to illustrate another type of heterogeneity, namely, that of wealth and income risks. That is, with forward planning under uncertainty, the risk environment of individuals or households becomes relevant. There can be individual shocks to income, such as a work layoff or a health problem, or aggregate shocks, such as an extended recession or stock market boom. Each of these shocks can differ in its duration – a temporary layoff can be usefully viewed as transitory, whereas a debilitating injury may affect income for many years. In planning consumption, it is important to understand the role of income risks and wealth risks. When there is no precautionary planning, such as when consumers have quadratic preferences, income risks do not become intertwined with other heterogeneous elements. However, when there is risk aversion, then the precise situation of individual income risks and insurance markets is relevant.

A commonly used model for income is to assume multiplicative permanent and transitory components, with aggregate and individual shocks, as in

$$\Delta \ln y_{it} = (\eta_t + \Delta u_t) + (\varepsilon_{it} + \Delta v_{it}).$$

Here $\eta_t + \Delta u_t$ is the common aggregate shock, with η_t a permanent component and Δu_t transitory. The idiosyncratic shock is $\varepsilon_{it} + \Delta v_{it}$, where ε_{it} is permanent and Δv_{it} transitory.

For studying individual level consumption with precautionary planning, it is standard

practice to assume constant relative risk aversion (CRRA) preferences and assume that the interest rate r_t is small. This, together with the income process above, gives a log-linear approximation to individual consumption growth

$$\Delta \ln c_{it} = \rho r_t + (\beta)\phi r_t)' z_{it} + k_1 \sigma_{At} + k_2 \sigma_{it} + \kappa_1 \eta_t + \kappa_2 \varepsilon_{it}. \quad (18)$$

Here, z_{it} reflects heterogeneity in preferences, such as differences in demographic characteristics. σ_{At} is the variance of aggregate risk and σ_{it} s the variance of idiosyncratic risk (with each conditional on what is known at time $t-1$), so that these terms reflect precautionary planning. Finally, η_t and ε_{it} arise because of adjustments that are made as permanent shocks are revealed. At time $t-1$ these shocks are not possible to forecast, but then they are incorporated in the consumption plan once they are revealed. In terms of the level of consumption c_{it}, Eq. (18) is written as

$$c_{it} = exp\left(\ln c_{it-1} + \rho r_t + (\beta + \phi r_t' z_{it} + k_1 \sigma_{At}\right) + k_2 \sigma_{it} + \kappa_1 \eta_t + \kappa_2 \varepsilon_{it}).$$

This is an intrinsically nonlinear model in the following heterogeneous elements: $\ln c_{it-1}$, z_{it},σ_{it} and ε_{it}. For aggregation, it seems we would need a great deal of distributional structure.

Here is where we can see the role of the risk environment, or markets for insurance for income risks. That is, if there were complete markets with insurance for all risks, then all risk terms vanish from consumption growth. When complete insurance exists for idiosyncratic risks only, then the idiosyncratic terms σ_{it} and ε_{it} vanish from consumption growth, since less precautionary saving is needed.

Otherwise, the idiosyncratic risk terms σ_{it} and ε_{it} represent heterogeneity that must be accommodated just like preference differences (and in other settings, participation differences).

In the realistic situation where risks are not perfectly insurable, we require distributional assumptions in order to formulate aggregate consumption. For instance, suppose that we assume that ($\ln c_{it\text{-}1}$; $(\beta + \phi r_t)' z_{it}$, ε_{it}) is joint normally distributed with $E_t(\varepsilon_{it}) = 0$, and that idiosyncratic risks are drawn from the same distribution for each consumer (so $\sigma_{it} = \sigma_{It}$ for each i), and that a stability assumption applies to the distribution of lagged consumption. Blundell and Stoker (2005) show that aggregate consumption growth is

$$\Delta \ln E_t(c_{it}) = \rho r_t + (\beta + \phi r_t)' E_t(z_{it}) + k_1 \sigma_{At} + k_2 \sigma_{It} + \kappa_1 \eta_t + \Lambda_t.$$

This model explains aggregate consumption growth in terms of the mean of preference heterogeneity, risk terms, and an aggregation factor Λ_t. The factor Λ_t is comprised of variances and covariances of the heterogeneous elements $\ln c_{it-1}$, z_{it} and ε_{it}. Thus, this model reflects how aggregate consumption will vary as the individual incomes become more or less risky, and captures how the income risk interplays with previous consumption values.

In overview, as micro consumption models are nonlinear, distributional restrictions are essential. On this point, an empirical fact is that the distribution of household consumption is often observed to be well approximated by a lognormal distribution, and so such lognormal restrictions may have empirical validity. Also relevant here is the empirical study of income and wealth risks, which has focused on earnings processes; see Meghir and Pistaferri (2004) for a recent contribution.

Micro to Macro and Vice Versa

We now turn to two related uses of aggregation structure that have emerged in the literature.

Aggregation as a Solution to Microeconometric Estimation

Consider a situation where the estimation of a model at the micro level is the primary goal of empirical work. Some recent work uses aggregation structure to enhance or permit micro-level parameter identification and estimation. Since aggregation structure provides a bridge between

models at the micro level and the aggregate level, it permits all data sources – individual-level data and aggregate-level data – to be used for identification and estimation of economic parameters. Sometimes it is necessary to combine all data sources to identify economic effects (for example, Jorgenson et al. 1982), and sometimes one can study (micro) economic effects with aggregate data alone (for example, Stoker 1986). Recent work has developed more systematic methods of using aggregate data to improve micro-level estimates. In particular, one can match aggregate data with simulated moments from the individual data as part of the estimation process.

To see how this can work, suppose we have data on labour participation over several time periods (or groups). We assume that the participation decision is given by the model (11) with normal unobserved heterogeneity, as discussed above. We normalize $\sigma_v = 1$ and take $s(t) = \psi$, a constant, so that the unknown parameters of the participation model are α,γ and ψ. The data situation is as follows; for each group $t = 1,...,T$, we observe the proportion of labour participants P_t and a random sample of benefits and schooling values, $\{B_{it}, S_{it}, i = 1,...,n_t\}$. Given the (probit) expression (13), estimation can be based on matching the observed proportion P_t to the simulated moment

$$\overline{P}_t(\alpha, \gamma, \psi) = \frac{1}{n_t} \sum_{i=1}^{n_t} \Phi[\psi - \alpha \ln B_{it} + \gamma \cdot S_{it}].$$

For instance, we could estimate by least squares over groups, by choosing $\hat{\alpha}, \hat{\gamma}, \hat{\psi}$ to minimize

$$\sum_{t=1}^{T} \left(P_t - \overline{P}_t(\alpha, \gamma, \psi)\right)^2.$$

Note that this approach does not require a specific assumption on the joint distribution of B_{it} and S_{it} for each t, as the random sample provides the distributional information needed to link the parameters to the observed proportion P_t.

It turns out that this approach for estimation is extremely rich, and was essentially mapped out by Imbens and Lancaster (1994). It has become a principal method of estimating demands for differentiated products, for use in structural models of industrial organization. See Berry et al. (2004) for good coverage of this development.

Can Macroeconomic Interaction Solve Aggregation Problems?

The basic heuristic that underlies much macroeconomic modelling is that, because of markets, individuals are very coordinated in their actions, so that individual heterogeneity likely has a secondary impact. In simplest terms, the notion is that common reactions across individuals will swamp any behavioural differences. This idea is either just wrong or, at best, very misleading for economic analysis. But that is not to deny that in real world economies there are many elements of commonality in reactions across individuals. Households face similar prices, interest rates and opportunities for employment. Extensive insurance markets effectively remove some individual differences in risk profiles. Optimal portfolio investment can have individuals choosing the same (efficient) basket of securities.

The question whether market interactions can minimize the impact of individual heterogeneity is a classic one, and by and large the answers are negative. However, there has been some recent work with calibrated stochastic growth models that raises some possibilities. A principal example of this is Krusell and Smith (1998), which we now discuss briefly. The Krusell–Smith set-up has infinitely lived consumers, with the same preferences within each period, but with different discount rates and wealth holdings. Each consumer has a chance of being unemployed each period, so there are transitory individual income shocks. Production arises from labour and capital, and there are transitory aggregate productivity shocks. Consumers can insure for the future by investing in capital only. Thus, insurance markets are incomplete, and consumers cannot hold negative capital amounts.

To make savings and portfolio decisions, consumers must predict future prices. To do this, each consumer must keep track of the evolution of the

entire distribution of wealth holdings, in principle. This is a lot of information to know, just like what is needed for standard aggregation solutions as discussed earlier. Krusell–Smith's simulations show, however, that this forecasting problem is much easier than one would suspect. That is, for consumer planning and for computing equilibrium, consumers get very close to optimal solutions by keeping track of only two things: mean wealth in the economy and the aggregate productivity shock. This is approximate aggregation, a substantial simplification of the information requirements that one would expect.

The source of this simplification, as well as its robustness, is a topic of active current study. One aspect is that most consumers, especially those with lowest discount rates, save enough to insure their risk so that their propensity to save out of wealth is essentially constant. Those consumers also hold a large fraction of the wealth, so that saving is essentially linear in wealth. This means that there is (approximate) exact aggregation structure, with the mean of wealth determining how much aggregate saving is undertaken. That is, the nature of savings and wealth accumulation approximately solves the aggregation problem for individual forecasting. Aggregate consumption, however, does not exhibit the same simplification. Many low-wealth consumers become unemployed and encounter liquidity constraints. Their consumption is much more sensitive to current output than that of wealthier consumers.

These results depend on the specific formulation of the growth model. Krusell and Smith (2006) survey work that suggests that their type of approximate aggregation can be obtained under a variety of variations of the basic model assumptions. As such, this work raises a number of fascinating issues on the interplay between economic interaction, aggregation and individual heterogeneity. However, it remains to be seen whether the structure of such calibrated models is empirically relevant to actual economies, or whether forecasting can be simplified even with observed variation in saving propensities of wealthy households.

Future Progress

Aggregation problems are among the most difficult in empirical economics. The progress that has been made recently is arguably due to two complementary developments. First is the enormous expansion in the availability of data on the behaviour of individual agents, including consumers, households, firms, and so on, in both repeated cross-section and panel data form. Second is the enormous expansion in computing power that facilitates the study of large data sources. These two trends can be reasonably expected to continue, which makes the prospects for further progress quite bright.

There is sufficient variety and complexity in the issues posed by aggregation that progress may arise from many approaches. For instance, we have noted how the possibility of approximate aggregation has arisen in computable stochastic growth models. For another instance, it is sometimes possible to derive properties of aggregate relationships with very weak assumptions on individual behaviour, as in Hildenbrand's (1994) work of the law of demand.

But is seems clear to me that the best prospects for progress lie with careful microeconomic modelling and empirical work. Such work is designed to ferret out economic effects in the presence of individual heterogeneity, and can also establish what are 'typical' patterns of heterogeneity in different applied contexts. Knowledge of typical patterns of heterogeneity is necessary for characterizing the distributional structure that will facilitates aggregation, and such distributional restrictions can then be refuted or validated with actual data. That is, enhanced understanding of the standard structure in the main application areas of empirical economics, such as with commodity demand, consumption and saving and labour supply, will lead naturally to an enhanced understanding of aggregation problems and accurate interpretation of aggregate relationships. There has been great progress of this kind in the past few decades, and there is no reason to think that such progress won't continue or accelerate.

Bibliography

Attanasio, O. 1999. Consumption. In *Handbook of macroeconomics*, ed. J. Taylor and M. Woodford, Vol. 1B. Amsterdam: North-Holland.

Banks, J., R. Blundell, and A. Lewbel. 1997. Quadratic Engel curves, indirect tax reform and welfare measurement. *Review of Economics and Statistics* 79: 527–539.

Berry, S., J. Levinsohn, and A. Pakes. 2004. Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of Political Economy* 112: 68–105.

Blundell, R. 1988. Consumer behaviour: Theory and empirical evidence. *Economic Journal* 98: 16–65.

Blundell, R., H. Reed, and T. Stoker. 2003. Interpreting aggregate wage growth. *American Economic Review* 93: 1114–1131.

Blundell, R., and T. Stoker. 2005. Aggregation and heterogeneity. *Journal of Economic Literature* 43: 347–391.

Browning, M., L. Hansen, and J. Heckman. 1999. Micro data and general equilibrium. In *Handbook of macroeconomics*, ed. J. Taylor and M. Woodford, Vol. 1A. Amsterdam: North-Holland.

Browning, M., and A. Lusardi. 1996. Household saving: Micro theories and micro facts. *Journal of Economic Literature* 34: 1797–1855.

Gorman, W.M. (Terence). 1953. Community preference fields. *Econometrica* 21: 63–80.

Hildenbrand, W. 1994. *Market demand: Theory and empirical evidence*. Princeton: Princeton University Press.

Imbens, G., and T. Lancaster. 1994. Combining micro and macro data in microeconometric models. *Review of Economic Studies* 61: 655–680.

Jorgenson, D., L. Lau, and T. Stoker. 1982. The transcendental logarithmic model of aggregate consumer behavior. In *Advances in econometrics*, ed. R. Basmann and G. Rhodes, Vol. 1. Greenwich: JAI Press.

Krusell, P., and A. Smith. 1998. Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy* 106: 867–896.

Krusell, P., and A. Smith. 2006. Quantitative macroeconomic models with heterogeneous agents. In *Advances in economics and econometrics, proceedings of the ninth world congress of the econometric society*, ed. R. Blundell, W. Newey, and T. Persson. Cambridge: Cambridge University Press.

Meghir, C., and L. Pistaferri. 2004. Income variance dynamics and heterogeneity. *Econometrica* 72: 1–32.

Stoker, T. 1986. Simple tests of distributional effects on macroeconomic equations. *Journal of Political Economy* 94: 763–795.

Stoker, T. 1993. Empirical approaches to the problem of aggregation over individuals. *Journal of Economic Literature* 31: 1827–1874.

Theil, H. 1954. *Linear aggregation of economic relations*. Amsterdam: North-Holland.

Aggregation (Production)

Jesus Felipe and Franklin M. Fisher

Abstract

Aggregation concerns the conditions under which several variables can be treated as one, or macro-relationships derived from micro-relationships. This problem is especially important in production, where, without proper aggregation, one cannot interpret the properties of the aggregate production function. The conditions under which aggregate production functions exist are so stringent that real economies surely do not satisfy them. The aggregation results pose insurmountable problems for theoretical and applied work in fields such as growth, labour or trade. They imply that intuitions based on micro variables and micro production functions will often be false when applied to aggregates.

Keywords

Aggregation (production); Cambridge capital theory debates; Capital aggregation; Cobb–Douglas functions; Endogenous growth; Growth accounting; Hicks, J.; Hicks–Leontief aggregation; Labour aggregation; Leontief, W.; National Income and Products Account (NIPA); Neoclassical growth theory; Output aggregation; Production functions; Productivity (measurement problems); Total factor productivity

JEL Classifications

E10; C43; B41; E01; E1; E23

Aggregation in production concerns the conditions under which macro production functions can be derived from micro production functions. Microeconomic theory elegantly treats the behaviour of optimizing individual agents in a world with an arbitrarily long list of individual commodities and prices. However, the desire to analyse the great

aggregates of macroeconomics – gross national product, inflation, unemployment, and so forth – leads to theories that treat such aggregates directly. The aggregation 'problem' matters because without proper aggregation one cannot interpret the properties of such macroeconomic models. This is particularly true as regards the production sector.

Leontief's Theorem

Underlying many results on aggregation is a theorem of Leontief (1947a, b). Let x and y be vectors of variables and $F(x, y)$ a twice-differentiable function. It is desired to aggregate over x, that is, to replace x with a scalar aggregator function, $g(x)$, such that $F(x) = H[g(x), y]$. This can be done if and only if, along any surface on which $F(x, y)$ is constant, the marginal rate of substitution between each pair of elements of x is independent of y. (For a proof, see Fisher 1993, pp. xiv–xvi.)

Hicks–Leontief Aggregation

Since optimizing, price-taking agents equate marginal rates of substitution to price ratios, one restriction permitting aggregation over commodities is the assumption that the prices of all goods to be included in an aggregate always vary proportionally.

This is called 'Hicks–Leontief aggregation' (Leontief 1936; Hicks 1939) and is a powerful expository tool. It requires no special assumptions as to the form of utility or production functions, but is applicable only in relatively artificial situations. Under more general circumstances, restrictions on utility or production functions become essential.

Aggregation in Consumption

Consider a single household. Suppose that we wish to describe behaviour in terms of aggregate commodities such as 'food' or 'clothing'. By Leontief's Theorem, a food aggregate exists if and only if the marginal rate of substitution between any two kinds of food is independent of consumption of any non-food commodity. If a similar restrictive condition is satisfied for all the aggregates to be constructed, then the household's utility function can be written in aggregate terms.

Even such restrictive conditions will not always suffice. If we wish to represent the household as maximizing the aggregate utility function subject to an aggregate budget constraint, we must have aggregate prices as well as aggregate consumption goods. This requires that aggregates such as 'food' be homothetic in their component variables, again considerably restricting the household's utility function (Gorman 1959; Blackorby et al. 1970).

Aggregation over agents presents a different set of questions. Suppose that we wish to treat the aggregate demands of a collection of households as the demands of a single, aggregate household. Then, only aggregate income and not its distribution can influence demand. At given prices, this makes the income derivative of every household's demand for a given commodity the same constant. Engel curves must be parallel straight lines. If zero income implies zero consumption, then all households must have the same homothetic utility function (Gorman 1953).

In general, the only consumer-theoretic restrictions obeyed by aggregate demand functions are those of continuity, homogeneity of degree zero, and the various restrictions implied by the budget constraint (cf. Sonnenschein 1972, 1973).

Aggregation in Production

A more detailed survey of much of what follows in this section is given in Felipe and Fisher (2003).

The analysis of aggregation conditions for production functions is far richer and the conditions even more demanding than in the case of demand functions.

Moreover, the subject has a complicated history and bears on the very foundations of neoclassical *macro*economics, negatively implicating the use of such important concepts as 'total factor productivity', 'natural rate of growth', 'capital–labour ratio', and even such terms as 'investment', 'capital', 'labour', and 'output'.

To take a simple example, suppose we have two production functions $Q^A = f^A\left(K_1^A, K_2^A, L^A\right)$ and $Q^B = f^B\left(K_1^B, K_2^B, L^B\right)$ for firms A and B, where $K_1 = K_1^A + K_1^B, K_2 = K_2^A + K_2^B$ and $L = L^A + L^B$ (K refers to capital – two types – and L to labour – assumed homogeneous). The problem is to determine whether and in what circumstances there exists a function $K = h(K_1, K_2)$ where the aggregator function $h(\cdot)$ has the property that $G(K, L) = G[h(K_1, K_2), L] = \Psi(Q^A, Q^B)$, and the function Ψ is the production possibility curve for the economy. Note that we have implicitly assumed that a production function exists for the firm. Further, even within the firm there is a problem of aggregation over factors. Here, we concentrate on aggregation over firms.

Klein (1946a, b) initiated the first debate on aggregation in production functions. He argued that the aggregate production function should be strictly a technical relationship, akin to the micro production function, and objected to utilizing the entire micro model with the assumption of profit-maximizing behaviour by producers in deriving the production functions of the macro model.

However, Kenneth May (1947) pointed out that this program is not generally achievable and, indeed, rests on a misunderstanding of what production functions actually are – even at the micro level. A production function does not tell us what outputs are or can be produced from a given set of inputs. It tells us what the *maximum* output is of a particular commodity, given a vector of inputs and the other outputs that are also to be produced from them.

That Klein's aggregation program is generally unachievable was specifically proved by André Nataf (1948). He showed that such aggregation is possible if and only if all micro production functions are additively separable in capital and labour.

The problem here is as follows. Suppose there are n firms indexed by $v = 1, \ldots, n$. Each produces the same output $Y(v)$ using the same type of labour $L(v)$, and a single type of capital $K(v)$. The vth firm has a two-factor production function $Y(v) = f^v\{K(v), L(v)\}$. The total output of the economy is $Y = \sum_v Y(v)$, total labour is $L = \sum_v L(v)$. Capital, on the other hand, may differ from firm to firm. Under what conditions can total output Y be written as $Y = \sum_v Y(v) = F(K, L)$ where $K = K\{K(1), \ldots, K(n)\}$ and $L = L\{L(1), \ldots, L(n)\}$ are indices of aggregate capital and labour, respectively? Nataf showed that, where the variables $K(v)$ and $L(v)$ are free to take on all values, the aggregate production function $Y = F(K, L)$ exists, if *and only if* every firm's production function is additively separable in labour and capital, that is, if every f^v can be written in the form $f^v\{K(v), L(v)\} = \phi^v\{K(v)\} + \psi^v\{L(v)\}$. Moreover, if one insists that labour aggregation be 'natural', with the L appearing in the aggregate production function, then all the $\psi^v\{L(v)\} = c\{L(v)\}$, where c is the same for all firms.

Nataf's theorem provides an extremely restrictive condition for inter-sectoral or even inter-firm aggregation. Evidently, aggregate production functions will not exist unless there are some further restrictions on the problem.

In fact, such restrictions are available; they stem from the requirement that a production function describe *efficient* production possibilities.

Capital Aggregation

Consider the simplest case of two factors, with physically homogeneous capital (K) and homogeneous labour (L), where total capital can be written as $K = \sum_v K(v)$, efficient production requires that aggregate output Y be maximized given aggregate labour (L) and aggregate capital (K). Under these simplified circumstances, it follows that $Y^M = F(K, L)$ where Y^M is maximized output, since, as was pointed out by May (1946, 1947), individual allocations of labour and capital to firms would be determined in the course of the maximization problem. This holds even if all firms have different production functions and whether or not there are constant returns.

In the (somewhat) more realistic case where only labour is homogeneous and technology is embodied in capital, Fisher (1965) proposed to treat the problem as one of labour being allocated to firms so as to maximize output, with capital

being firm-specific. Here, no 'natural' aggregate of capital exists.

Given that output is maximized with respect to the allocation of labour to firms, with such maximized output denoted by Y^*, the question becomes: under what circumstances is it possible to write total output as $Y^* = F(J, L)$ where $J = J\{K(1), \ldots, K(n)\}$, where $K(v)$, $v = 1, \ldots, n$, represents the stock of capital of each firm (that is, one kind of capital per firm)? Since the values of $L(v)$ are determined in the optimization process there is no labour aggregation problem. The entire problem in this case lies in the existence of a capital aggregate. Since Leontief's condition is both necessary and sufficient for the existence of a group capital index, the previous expression for Y^* is equivalent to $Y^* = G\{K(1), \ldots, K(n), L\}$ if and only if the marginal rate of substitution between any pair of the $K(v)$is independent of L.

Fisher drew the implications of this condition. He showed that, under strictly diminishing returns to labour ($f^v_{LL} < 0$), if any one firm has an additively separable production function (that is, $f^v_{LL} \equiv 0$), then a necessary and sufficient condition for capital aggregation is that *every* firm have such a production function. (Throughout, such subscripts denote partial differentiation in the obvious manner.) This means that capital aggregation is impossible if there is both a firm which uses labour and capital in the same production process, and another one which has a fully automated plant. Fisher found that a necessary and sufficient condition for capital aggregation is that every firm's production function satisfy a partial differential equation in the form $f^v_{KL}/f^v_K\, f^v_{LL} = g\left(f^v_L\right)$, where g is the same function for all firms. More important, on the assumption of constant returns to scale, the case of capital-augmenting technical differences (that is, embodiment of new technology can be written as the product of the amount of capital times a coefficient) turns out to be the *only case* in which a capital aggregate exists. This means that each firm's production function must be writeable as $F(b_v\, K_v, L_v)$, where the function $F(\cdot,\cdot)$ is common to all firms, but the parameter b_v can differ. Under these circumstances, a unit of one type of new capital equipment is the exact duplicate of a fixed number of units of old capital equipment ('better' is equivalent to 'more'). As we would expect, given constant returns to scale, the aggregate stock of capital can be constructed with capital measured in efficiency units. Fisher (1965) could not come up with a closed-form characterization of the class of cases in which an aggregate stock of capital exists when the assumption of constant returns is dropped. Nevertheless, as he showed, there do exist classes of non-constant returns production functions which do allow construction of an aggregate capital stock. On the other hand, if constant returns are not assumed there is no reason why perfectly well-behaved production functions cannot fail to satisfy Fisher's partial differential equation given above. Capital aggregation is then impossible if any firm has one of these 'bad apple' production functions. To sum up: aggregate production functions exist if and only if all micro production functions are identical except for the capital efficiency coefficient – an extremely restrictive condition.

Working with the profits function rather than with the production function, Gorman (1968) reached similar conclusions to those of Fisher.

Fisher extended his original work. First of all, he analysed (Fisher 1965) the case where each firm produces a single output with a single type of labour, but two capital goods, that is, $Y(v) = f^v(K_1, K_2, L)$. Here Fisher distinguished between two different cases. The first is that of aggregation across firms over one type of capital (for example, plant or equipment). Fisher concluded that the construction of a sub-aggregate of capital goods requires even more stringent conditions than for the construction of a single aggregate. For example, if there are constant returns in K_1, K_2 and L, there will not be constant returns in K_1 and L, so that the difficulties of the two-factor non-constant returns case appear. Further, if the vth firm has a production function with all three factors as complements, then no K_1 aggregate can exist. Thus, for example, if any firm has a generalized Cobb–Douglas production function (with the v argument omitted) in plant, equipment, and labour $Y = AK_1^{\alpha}K_2^{\beta}L^{1-\alpha-\beta}$, one cannot construct a separate plant or separate equipment aggregate for the

economy as a whole (although this does not prevent the construction of a full capital aggregate).

The other case Fisher (1965) considered was that of the construction of a complete capital aggregate. In this case, a necessary condition is that it be possible to construct such a capital aggregate for each firm taken separately; and a necessary and sufficient condition (with constant returns), given the existence of individual firm aggregates, is that all firms differ by at most a capital augmenting technical difference. They can differ *only* in the way in which their individual capital aggregate is constructed.

Second, Fisher (1982) asked whether the crux of the aggregation problem derives from the fact that capital is considered to be an immobile factor. He showed that the aggregation problem seems to be due only to the fact that capital is fixed and is not allocated efficiently. That is true in the context of a two-factor production function. However, if one works in terms of many factors, all mobile over firms, and asks when it is possible to aggregate them into macro groups, the mobility of capital has little bearing on the issue. In fact, where there are several factors, each of which is homogeneous, optimal allocation across *firms* does not guarantee aggregation across *factors*. The conditions for the existence of such aggregates are still very stringent, but this has to do with the necessity of aggregating over firms rather than with the immobility of capital. A possible way of interpreting the existence of aggregates at the firm level is that each firm could be regarded as having a two-stage production process. In the first one, the factors to be aggregated, $X_i(v)$, are combined to produce an intermediate output, $\phi^v(X(v))$. This intermediate output is then combined with the other factor, $L(v)$, to produce the final output. Aggregation of X can be done if and only if firms are either all alike as regards the first stage of production, or all alike as regards the second stage. If they are all alike as regards the first stage, then the fact that L is mobile plays no role. If they are all alike as regards the second stage, then the fact that the X_i are mobile plays no role.

Finally, Fisher (1983) is another extension of the original problem to study the conditions under which full and partial capital aggregates, such as 'plant' or 'equipment', would exist simultaneously. Not surprisingly, the results are as restrictive as those above. See also Blackorby and Schworm (1984).

Labour and Output Aggregation

Fisher (1968) went on to study the problems involved in labour and output aggregation, pointing out that the aggregation problem is not restricted to capital. Output aggregation and labour aggregation are also necessary if one wants to use a sector-wide or economy-wide aggregate production function.

Fisher again studied aggregation over firms, with labours and outputs shifted over firms to achieve efficient production, given the capital stocks. In the simplest case of constant returns, a labour aggregate will exist if and only if a given set of relative wages induces all firms to employ different labours in the same proportions. Similarly, where there are many outputs, an output aggregate will exist if and only if a given set of relative output prices induces all firms to produce all outputs in the same proportion. Thus, the existence of a labour aggregate requires the absence of specialization in employment; and the existence of an output aggregate requires the absence of specialization in production – indeed, all firms must produce the same market basket of outputs differing only in their scale. (Blackorby and Schworm 1988, is an extension of Fisher 1968.)

Houthakker–Sato Aggregation Conditions

Whereas Fisher sought to develop conditions where aggregate production functions would always work, Houthakker (1955–56) and Sato (1975) considered two-factor cases in which the problem was restricted by assuming that the distribution of capital over firms remains constant. In such cases it is obvious that one can aggregate over capital. Houthakker and Sato's contributions (see also Levhari 1968) were to show the relationships between the fixed distribution of capital and the form of the aggregate production function.

Fisher's Simulations

But, if aggregate production functions do not exist, how is it that they appear to 'work' in the sense that they fit the data well, that the estimated elasticities are close to the factor shares, and that wage rates are approximate the calculated marginal product of labour? We shall have more to say on this below, but here consider another result of Fisher (1971). This paper reports the results of simulations in a simple (heterogeneous capital, homogeneous labour and output) economy in which the aggregation conditions are known not to be satisfied. The principal result is that when, despite this, calculated factor shares just happen to be roughly constant, then the Cobb–Douglas aggregate production function 'works' in the above sense, even though the approximate constancy of factor shares *cannot* be caused by the non-existent aggregate production function. (See Fisher et al. 1977 for the case of the CES production function.)

Implications for Empirical Work

Empirically, the non-existence of the aggregate production function poses a conundrum. If aggregate production functions do not exist, there must be some other reason why they seem to work empirically. The answer has been in the literature for a long time (Simon and Levy 1963; Simon 1979; Shaikh 1980), and more recently Felipe (2001) and Felipe and McCombie (2001, 2002, 2003, 2005, 2006a, b) have elaborated upon it. (For an in-depth discussion of these issues see the papers in the *Eastern Economic Journal* 2005.) However, like the theoretical arguments underlying the non-existence of the aggregate production function, these arguments have largely been ignored.

The argument is that, because the data used in aggregate empirical applications are not physical quantities but values, the accounting identity that relates definitionally the value of total output to the sum of the value of total inputs can be rewritten as a form that resembles a production function.

More specifically, the National Income and Products Account (NIPA) identity states that value added equals the wage bill plus total profits, that is,

$$V_t \equiv W_t + \Pi_t \equiv w_t L_t + r_t J_t \qquad (1)$$

where V is real value added, W is the total wage bill in real terms, Π denotes total profits ('operating surplus', in the NIPA terminology), also in real terms, w is the average real wage rate, L is employment, r is the average *ex post* real profit rate, and J is the deflated or constant-price value of the stock of capital. (Expression (1) is an accounting identity, not the result of Euler's Theorem.) In applied aggregate work, the measures of output and capital used are the constant-price values, not physical quantities. We denote them by V and J, respectively. These are different from Y and K used above, which denoted physical quantities. The symbol $\equiv$ indicates that expression (1) is an accounting identity.

Expressing the identity (1) in growth rates yields:

$$\hat{V}_t \equiv a_t \hat{w}_t + (1 - a_t)\hat{r}_t + a_t \hat{L}_t + (1 - a_t)\hat{J}_t \qquad (2)$$

where ^ denotes a proportional growth rate, $a_t \equiv w_t L_t / V_t$ is the share of labour in output, and $1 - a_t \equiv r_t J_t / V_t$ is the share of capital. So far no assumption of any kind has been made.

Suppose now that factor shares in the economy are relatively stable. This could be due, for example, to the fact that firms set prices according to a mark-up on unit labour costs. Assume also that w_t and r_t grow at constant rates. Then

$$\hat{V}_t \equiv \lambda + a\hat{L}_t + (1 - a)\hat{J}_t \qquad (3)$$

Where $\lambda \equiv a\hat{w} + (1 - a)\hat{r}$. Integrating (3) and taking antilogarithms,

$$V_t \equiv A_0 \exp(\lambda t) L_t^a J_t^{1-a} \qquad (4)$$

Expression (4) is simply the NIPA accounting identity, expression (1), rewritten under the two assumptions mentioned above. It is certainly not a Cobb–Douglas production function, as such does not exist.

What are the implications of this argument? Suppose one estimates the standard

Cobb–Douglas regression $V_t \equiv C_0 \exp(\gamma t)\ L_t^{\alpha_1} J_t^{\alpha_2}$ and in this economy factor shares are approximately constant and wage and profit rate growth is approximately constant. Then, this regression will yield very good results, since it approximates the identity (4). The statistical fit will be close to unity, $\alpha_1 \cong a$, $\alpha_2 \cong 1 - a$, and $\gamma \cong \lambda$. However, the aggregate production function may not exist, or firms in this economy may be subject to increasing returns to scale, although the regression results might lead us to believe otherwise.

On the other hand, if the assumptions about the path of the factor shares and the growth rates of w and r are incorrect, the regression $V_t \equiv C_0 \exp(\gamma t) L_t^{\alpha_1} J_t^{\alpha_2}$ will not yield good results. Felipe and Holz 2001, showed using Monte Carlos simulations that the main reason why the Cobb–Douglas regression $V_t \equiv C_0 \exp(\gamma t) L_t^{\alpha_1} J_t^{\alpha_2}$ often fails is that the approximation of $[a_t \hat{w}_t + (1 - a_t)\hat{r}_t]$ through the constant term λ is incorrect. Such widely discussed problems as unit roots or endogeneity of the regressors are not the key issues. This simply means that we have to search for better approximations to the identity. (See Felipe and McCombie 2001, 2003, for the derivations of the CES and translog approximations to the accounting identity.)

These results have devastating implications for empirical neoclassical macro growth theory, including endogenous growth, and total factor productivity measurement and growth accounting exercises. Indeed, Felipe and McCombie (2006b) have shown using simulations that the true rate of technical progress, computed with the use of firm-level data, is very different from that obtained with the use of aggregate data. Indeed, the two measures of productivity are so far apart that it is concluded that total factor productivity growth calculated with aggregate data is in no way a proxy for the true rate of technological progress.

Why Do Economists Continue Using Aggregate Production Functions?

Most economists are not aware of these results, but simply think of the aggregate production function as part of their basic toolkit. Others use such concepts as total productivity growth without realizing that they are assuming the existence of a non-existent construct.

Some economists, on the other hand, are aware of the aggregation results and yet continue using aggregate production functions. The reasons for doing so fall under three broad categories:

1. Aggregate production functions are seen as useful parables (Samuelson 1961–62).
2. So long as aggregate production functions appear to give empirically reasonable results, why shouldn't they be used?
3. For the applications where aggregate production functions are used, there is no other choice.

However, in the light of the aggregation results, none of these reasons seems valid. Samuelson's parable argument was stated in the context of the so-called Cambridge capital theory debates. (It should not be thought that the aggregation problems have no bearing on the Cambridge–Cambridge debates. The discovery that aggregate production functions can violate properties that one expects of production functions, so-called reswitching and reverse capital-deepening, was at bottom a discovery that the aggregate concept used is not a production function at all. The aggregation problem literature shows that this was to be expected.) Samuelson showed that even in cases with heterogeneous capital goods some rationalization could be provided for the validity of the neoclassical parable, which assumes that there is a single homogenous factor referred to as capital, whose marginal product equals the interest rate. But Samuelson's results hold only in very restrictive cases, as we should expect from the aggregation literature. (See also Garegnani 1970.)

A variation of the parable argument is that the aggregate production function should be understood as an *approximation*. It is evident that Fisher's (exact) aggregation conditions are so stringent that one can hardly believe that actual economies will satisfy them even approximately. Fisher (1969), therefore, asked: What about the possibility of a *satisfactory approximation*? Thus,

suppose the values of capitals and labours in the economy lie in a bounded set and the requirement is that an aggregate production function lie within some specified distance of the true production surface for all points in the bounded set. Can this happen without the approximate satisfaction of the aggregation conditions? Fisher showed that this cannot reasonably happen by proving that the *only* way for approximate aggregation to hold without approximate satisfaction of the Leontief conditions is for the derivatives of the functions involved to wiggle violently up and down, an unnatural property not exhibited by the aggregate production functions used in practice.

The second argument is that, despite the aggregation results, neoclassical macroeconomic theory generally deals with macroeconomic aggregates derived by analogy with the micro concepts. Then, the argument goes, why not continue using them? Naturally, the aggregation problem appears in all areas of economics, including consumption theory, where a well-defined micro consumption theory exists. The neoclassical aggregate production function is also built by analogy (Ferguson 1971).

This argument is untenable. Employing macroeconomic production functions on the unverified premise that inference by analogy is correct is inadmissible. Further, as opposed to the (already suspect) case of the consumption function, the conditions for successful aggregation of production functions seem far more outlandish.

The third and final argument given for the use of aggregate production functions is that there is no other option if one is to answer the questions for which the aggregate production function is used, for example to discuss productivity differences across nations. But, 'It's crooked, but it's the only wheel in town' is not a scientific argument. The profession needs to find a different 'wheel'.

See Also

- ▶ Aggregation (Theory)
- ▶ Cost Functions
- ▶ Endogenous Growth Theory
- ▶ Growth Accounting
- ▶ Neoclassical Growth Theory
- ▶ Production Functions
- ▶ Total Factor Productivity

Bibliography

Blackorby, C., and W. Schworm. 1984. The structure of economies with aggregate measures of capital: A complete characterization. *Review of Economic Studies* 51: 633–650.

Blackorby, C., and W. Schworm. 1988. The existence of input and output aggregates in aggregate production functions. *Econometrica* 56: 613–643.

Blackorby, C., G. Lay, D. Nissen, and R. Russell. 1970. Homothetic separability and consumer budgeting. *Econometrica* 38: 468–472.

Eastern Economic Journal. 2005. Symposium on the aggregate production function. 31(3).

Felipe, J. 2001. Endogenous growth, increasing returns, and externalities: An alternative interpretation of the evidence. *Metroeconomica* 52: 391–427.

Felipe, J., and F. Fisher. 2003. Aggregation in production functions: What applied economists should know. *Metroeconomica* 54: 208–262.

Felipe, J., and C. Holz. 2001. Why do aggregate production functions work? Fisher's simulations, Shaikh's identity, and some new results. *International Review of Applied Economics* 15: 261–285.

Felipe, J., and J. McCombie. 2001. The CES production function, the accounting identity, and Occam's razor. *Applied Economics* 33: 1221–1232.

Felipe, J., and J. McCombie. 2002. A problem with some recent estimations and interpretations of the mark-up in manufacturing industry. *International Review of Applied Economics* 16: 187–215.

Felipe, J., and J. McCombie. 2003. Some methodological problems with the neoclassical analysis of the East Asian miracle. *Cambridge Journal of Economics* 54: 695–721.

Felipe, J., and J. McCombie. 2005. Why are some countries richer than others? A skeptical view of Mankiw–Romer–Weil's test of the neoclassical growth model. *Metroeconomica* 56: 360–392.

Felipe, J. and McCombie, J. 2006a. Is a theory of total factor productivity really needed? *Metroeconomica* (forthcoming).

Felipe, J. and McCombie, J. 2006b. The tyranny of the identity: Growth accounting revisited. *International Review of Applied Economics* (forthcoming).

Ferguson, C. 1971. Capital theory up to date: A comment on Mrs. Robinson's article. *Canadian Journal of Economics* 4: 250–254.

Fisher, F. 1965. Embodied technical change and the existence of an aggregate capital stock. *Review of Economic Studies* 32: 263–288.

Fisher, F. 1968. Embodied technology and the existence of labor and output aggregates. *Review of Economic Studies* 35: 391–412.
Fisher, F. 1969. Approximate aggregation and the Leontief conditions. *Econometrica* 37: 457–469.
Fisher, F. 1971. Aggregate production functions and the explanation of wages: A simulation experiment. *Review of Economics and Statistics* 53: 305–325.
Fisher, F. 1982. Aggregate production functions revisited: The mobility of capital and the rigidity of thought. *Review of Economic Studies* 49: 615–626.
Fisher, F. 1983. On the simultaneous existence of full and partial capital aggregates. *Review of Economic Studies* 50: 197–208.
Fisher, F. 1993. *Aggregation: Aggregate production functions and related topics*. Cambridge, MA: MIT Press.
Fisher, F., R. Solow, and J. Kearl. 1977. Aggregate production functions: Some CES experiments. *Review of Economic Studies* 44: 305–320.
Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37: 407–436.
Gorman, W. 1953. Community preference fields. *Econometrica* 21: 63–80.
Gorman, W. 1959. Separable utility and aggregation. *Econometrica* 27: 469–481.
Gorman, W. 1968. Measuring the quantities of fixed factors. In *Value, capital, and growth: Papers in honour of Sir John Hicks*, ed. J. Wolfe. Edinburgh: Edinburgh University Press.
Hicks, J. 1939. *Value and capital*. Oxford: Clarendon Press.
Houthakker, H. 1955. The Pareto distribution and the Cobb–Douglas production function in activity analysis. *Review of Economic Studies* 23: 27–31.
Klein, L. 1946a. Macroeconomics and the theory of rational behavior. *Econometrica* 14: 93–108.
Klein, L. 1946b. Remarks on the theory of aggregation. *Econometrica* 14: 303–312.
Leontief, W. 1936. Composite commodities and the problem of index numbers. *Econometrica* 4: 39–59.
Leontief, W. 1947a. Introduction to a theory of the internal structure of functional relationships. *Econometrica* 15: 361–373.
Leontief, W. 1947b. A note on the interrelationship of subsets of independent variables of a continuous function with continuous first derivatives. *Bulletin of the American Mathematical Society* 53: 343–350.
Levhari, D. 1968. A note on Houthakker's aggregate production function in a multi-firm industry. *Econometrica* 36: 151–154.
May, K. 1946. The aggregation problem for a one-industry model. *Econometrica* 14: 285–298.
May, K. 1947. Technological change and aggregation. *Econometrica* 15: 51–63.
Nataf, A. 1948. Sur la possibilité de construction de certains macromodč˘es. *Econometrica* 16: 232–244.
Samuelson, P. 1961. Parable and realism in capital theory: The surrogate production function. *Review of Economic Studies* 29: 193–206.
Sato, K. 1975. *Production functions and aggregation*. Amsterdam: North-Holland.
Shaikh, A. 1980. Laws of production and laws of algebra: Humbug II. In *Growth, profits and property: Essays in the revival of political economy*, ed. E. Nell. Cambridge: Cambridge University Press.
Simon, H. 1979. On parsimonious explanations of production relations. *Scandinavian Journal of Economics* 81: 459–474.
Simon, H., and F. Levy. 1963. A note on the Cobb–Douglas function. *Review of Economic Studies*: 93–94.
Sonnenschein, H. 1972. Market excess demand functions. *Econometrica* 40: 549–563.
Sonnenschein, H. 1973. Do Walras' identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6: 345–354.

Aggregation (Theory)

Werner Hildenbrand

Abstract

The aim of aggregation theory is to link the micro and macroeconomic notions of aggregate demand. One would like such a link to exist for any heterogeneous population, for a large set of all conceivable income assignments, and for a small number of statistics of the income distribution. This cannot be achieved. What can be achieved is critically discussed in section "Income Aggregation". In section "Monotone Mean Demand", another important topic of aggregation theory is considered: how does mean demand react to price changes? As an example, the 'law of demand' is discussed.

Keywords

Aggregate demand; Aggregation; Behavioural heterogeneity; Exact income aggregation; Law of demand; Monotonicity; Revealed preferences; Slutzky substitution effect

JEL classifications

C43

Introduction

Aggregation theory of demand aims at identifying observable explanatory variables for aggregate demand starting from a microeconomic description of the underlying population of households. In the simple case, where the demand decision of a household is the choice of a commodity vector in a budget set, which is determined by the price vector p and income x (total expenditure), the demand behaviour of a household h is modelled by a demand function $f^h(p,x) \in \mathbb{R}^l_+$ (commodity space), which is defined for every strictly positive price vector $p \in P$ and every income level $x \geq 0$. The demand function f^h might, but need not be derived from preference maximization under the budget constraint.

Aggregate demand is defined as *mean demand across the population H*, that is to say, $\frac{1}{H}\sum_{h\in H} f^h\left(p,x^h\right)$. The population H is viewed as heterogeneous in income and demand behaviour. Thus, mean demand is determined by the price vector p and the joint distribution of income x^h and demand function f^h across the population H.

This general microeconomic definition of mean demand is sufficiently specific for certain problems in pure theory, for example for the existence problem in general equilibrium theory.

In macroeconomics or in applied demand analysis the notion of aggregate demand is quite different. There the explanatory variables for aggregate demand are the price vector and certain statistics $S(G_x)$ of the income distribution function G_x such as mean income, a measure of income inequality (for example, the variance of log income) or higher moments of the income distribution. In any case, no household specific variable is used in the aggregate demand function. The aim of the aggregation theory is to link the micro and macroeconomic notions of aggregate demand. More specifically, given an assignment $(f^h)_{h\in H}$ of demand functions and a set $X \subset \mathbb{R}^H_+$ of income assignments $(x^h)_{h\in H}$, one seeks for a representation of mean demand of the following form: there exists a function F from $P \times \mathbb{R}^N$ into $\mathbb{R}^l_+$ and N statistics $S_1(G_x)$, … , $S_N(G_x)$ of the income distribution function G_x, such that

$$\frac{1}{H}\sum_{h\in H} f^h\left(p,x^h\right) = F(p,S_1(G_x),\ldots,S_N(G_x)). \tag{1}$$

for all income assignments $(x^h)_{h\in H}$ in X *and all* price vectors p in P.

One would like such a representation to exist for any heterogeneous population H, for a large set X, ideally for all conceivable income assignments, that is, $X = \mathbb{R}^H_+$ and for a small number N of statistics. This, of course, cannot be achieved.

The theory of income aggregation is surveyed in section Income Aggregation, where also basic references are given. The main results are:

- A representation of the form (1), which must hold in the case $X = \mathbb{R}^H_+$ is an unreasonable strong requirement. Indeed, if a representation exists, then the population H must be homogeneous in demand behaviour, that is, $f^h = f$ for all $h \in H$, and furthermore
- If N is less than the number of households in H and the common demand function f has the basic properties of demand theory (budget identity and homogeneity), then either f is linear in income or at least for one commodity i, the income share function $w_i(p, x) := p_i f_i(p, x)/x$ is oscillating (that is, the derivative $\partial_x w_i(p, \cdot)$ changes its sign infinitely often).

Thus, households' behaviour which is modelled by the common demand function is either unreasonably simple or incredibly sophisticated. These results clearly show that the requirement $X = \mathbb{R}^H_+$ leads to an ill-posed problem.

For a heterogeneous population H there exists (see Example 3) a finite partition $\{X_k\}_{k\in K}$ of the set $\mathbb{R}^H_+$ of all conceivable income assignments and for every $k \in K$ there is a function $F^k(p, G)$, where G denotes an income distribution function, such that

$$\frac{1}{H}\sum_{h\in H} f^h\left(p,x^h\right) = F^k(p,G_x) \tag{2}$$

for every income assignment $(x^h)_{h\in H}$ in the set X^k and for every $p \in P$.

Thus, for a heterogeneous population H, there is no closed-form definition of an aggregate demand function; there is only a piecewise one, since the aggregate demand functions F^k and F^j are different for $k \neq j$. The less heterogeneous the population the coarser the partition, that is, the smaller is $\#K$. The sets X^k of the partition are large (see Example 3), in particular, if $\left(x_0^h\right) \in X^k$, then for every strictly increasing function ϕ the income assignment $x^h = \phi\left(x_0^h\right), h \in H$, also belongs to X^k (see Figs. 3 and 4).

The aggregate demand functions $F^k(p, G)$ in (2) require the knowledge of the entire income distribution. In many applications one might assume that the distribution of relevant income assignments in the set X^k can be modelled by some few parameters (structural stability of income distributions). For example, if the population is 'very large' one might restrict attention to those (x^h) in X^k whose distributions are (approximately) log normal. Then, on this subset of X^k, mean demand has a representation of the form $F^k(p,\bar{x},\sigma)$, where $\bar{x}$ denotes mean income across the population and σ^2 is the variance of log income, which can be interpreted as a measure of income inequality.

Another important topic of aggregation theory is to analyse how mean demand of a heterogeneous population reacts to price changes under the ceteris paribus clause that households' income and demand functions remain fixed. In this case mean demand is denoted by $F(p)$. Among the various desirable dependence structures is certainly the 'law' of demand, which asserts that the vector $\Delta p \in \mathbb{R}^l$ of price changes and the resulting vector $\Delta p \in \mathbb{R}^l$ of mean demand changes point in opposite directions, that is, the scalar product $\Delta p \cdot \Delta F := \sum_{i=1}^{l} \Delta p_i \Delta F_i$ is negative.

Certainly, the 'law' is not meant to be an empirical law, but a monotonicity property of the mean demand function $F(p)$ which is defined under a ceteris paribus clause in a mathematical model of a population of households. Thus, the 'law' asserts that the mean demand function F is *strictly monotone*, that is,

$$(p-q)\cdot(F(p)-F(q)) < 0 \text{ for all } p \neq q \text{ in } P$$

In particular, every partial mean demand curve is strictly decreasing. This partial monotonicity property, however, is not sufficient for proving the uniqueness and stability of the equilibria for a multi-commodity demand-supply system; one needs strict monotonicity in the multi-commodity version.

Which behavioural assumption on the household level and/or which form of heterogeneity of the population lead to monotone mean demand? To answer this question one assumes that demand functions f^h satisfy the weak axiom of revealed preferences or, more specifically, that they are derived from preference maximization. Then, partial monotonicity is easily obtained, for example, by excluding inferior goods. However, multi-commodity monotonicity is more difficult to obtain. Trivially, mean demand is monotone if all demand functions $f^h(p, x^h)$ were monotone in p. This, however, requires that either $f^h(p, \cdot)$ is linear in income or that the Slutzky substitution effect is sufficiently strong. (For a precise formulation, see the Theorem of Mitjuschin and Polterovich 1978; law of demand.) Since the Slutzky substitution effect might be arbitrarily small, one is interested in finding alternative assumptions, which do not rely on a strong Slutzky substitution effect. These assumptions should not require that households' demand functions are monotone. Obviously, to obtain the desirable aggregation effect, the population must be heterogeneous. Thus, in contrast to the problem of income aggregation, heterogeneity does not complicate the analysis, yet it is necessary to obtain monotonicity of mean demand by aggregation. More details are given in section Monotone Mean Demand. For example, let H be a population which is homogeneous in demand behaviour, that is, $f^h = f$, $h \in H$ and the common demand function is not monotone. However, the population is heterogeneous in income.

Then, for a given income assignment $(x^h)_{h \in H}$, mean demand $F^H(p)$ is not monotone in p. If one increases now the population size, that is, the number $\#H$ of households tends to infinity and if for increasing $\#H$ the income distribution functions G^H of households in H converge to a concave distribution function G, then, for $\#H$ sufficiently large, mean demand $F^H(p)$ is

'approximately' monotone, that is to say, $F^H(p)$ converges to a monotone function. Consequently, in the limit, that is, for an indefinitely large population which admits a concave income distribution function, mean demand is monotone. The mathematical model for such a limit population cannot be a finite or countably infinite set; it must be an atomless measure space of households, for example, the unit interval [0,1] with Lebesgue measure (continuum of households).

If these large populations are heterogeneous in income and demand behaviour, then one can meaningfully pose the problem of 'smoothing by aggregation': is mean demand continuous or differentiable without assuming these properties on the household level? The basic reference is Trockel (1984).

Finally, one should mention the literature on 'behavioural heterogeneity' initiated by Grandmont (1992). Here the goal is to obtain a stronger property than strict monotonicity of mean demand: diagonal dominance of the Jacobian $\partial_p F(p)$ of mean demand in the sense that

$$p_i \partial_{p_i} F_i(p) > \sum_{j \neq i} p_j \left| \partial_{p_j} F_i(p) \right|$$

and

$$p_i \partial_{p_i} F_i(p) > \sum_{j \neq i} p_j \left| \partial_{p_i} F_j(p) \right|.$$

This diagonal dominance models a strong restriction on the interdependence among the various commodity markets and is the basis for partial equilibrium analysis. For a general discussion of 'behavioural heterogeneity' see Hildenbrand and Kneip (2005).

Income Aggregation

The demand behaviour of every household h in a population H is modelled by a demand function $f^h \in \mathcal{F}$. In this section it is not required that demand functions are derived by preference maximization under budget constraints. One only needs that demand functions $f^h \in \mathcal{F}$ are continuous functions from $P \times \mathbb{R}_+$ into $\mathbb{R}^l_+$ with $f(p, 0) = 0$, where P denotes the set of all strictly positive price vectors in $\mathbb{R}^l$.

For every income assignment $(x^h)_{h \in H}$, $x^h \geq 0$, we consider *mean demand* $\frac{1}{H} \sum_{h \in H} f^h(p, x^h)$. The 'problem of income aggregation' has been defined in the literature by the qst: does there exist a function F from $P \times \mathbb{R}_+$ into $\mathbb{R}^l_+$ such that

$$\frac{1}{H} \sum_{h \in H} f^h(p, x^h) = F(p, \bar{x}), \text{where } \bar{x} = \frac{1}{H} \sum_{H} x^h, \tag{3}$$

for all income assignments in a given set $X \subset \mathbb{R}^H_+$ and all $p \in P$?

If one asks this question for all conceivable income assignments, that is, $X = \mathbb{R}^H_+$, then this is an ill-posed problem since it allows only a trivial solution.

Theorem (Antonelli 1886): *There exists a function $F(p, \bar{x})$ such that (3) holds on $\mathbb{R}^H_+ \times P$ if and only if the population H is homogeneous in demand behaviour, that is, $f^h = f$, and furthermore $f(p, x)$ is linear in x, that is, $f(p, x) = \alpha(p)x$, $\alpha(p) \in \mathbb{R}^l_+$ Thus, $F(p, \bar{x}) = \alpha(p)\bar{x}$.*

One might ask whether a less restrictive condition than (3) allows for a nontrivial solution. That is to say, one might consider mean demand functions that depend on a wider set of aggregate income variables than just mean income, for example, the variance or higher moments of the distribution of income. The answer is definitely negative.

For every income assignment $(x^h)_{h \in H}$, let G_x denote its distribution function, that is,

$$G_x(\xi) := \frac{1}{H} \left\{ h \in H \,|\, x^h \leq \xi \right\}, \xi \in \mathbb{R}$$

Proposition 1 *There exists a function $F(p, G)$ such that*

$$\frac{1}{H} \sum_{h \in H} f^h(p, x^h) = F(p, G_x) \tag{4}$$

for all conceivable income assignments, that is, $X = \mathbb{R}_+^H$ and all $p \in P$, if and only if the population H is homogeneous in demand behaviour, that is, all households in H have the same demand function. Then $F(p, G_x) = \int f(p, \xi)\, dG_x(\xi)$.

Proof Consider any two households k and j in H, and an income assignment $(x^h)_{h \in H}$ with $x^k > 0$ and $x^j = 0$. Now one interchanges the income of households k and j. This does not change the distribution function of income. Hence property (4) and the fact that $f^k(p, 0) = f^j(p, 0) = 0$ implies that $f^k(p, x^k) = f^j(p, x^k)$. Since this holds for all $x^k > 0$ and $p \in P$ one obtains $f^k = f^j$. On the other hand, if $f^h = f$ for all $h \in H$, then $\frac{1}{H} \sum_{h \in H} f^h(p, x^h) = \int f(p, x) dG_x =: F(p, G_x)$.

The justification for considering the generalized problem of income aggregation as defined by (4) is based on the view that for large populations, which this survey emphasizes, income distribution functions can often be modelled by some few parameters, for example, log-normal distributions.

By Proposition 1 it is clear that one is forced to restrict the set X of admissible income assignments if one wants to escape the case of trivial solutions, $f^h = f$, to the aggregation problem as defined by (4). Motivated by the special role which zero income and the assumption $f(p, 0) = 0$ play in the proofs of Antonelli's Theorem or Proposition 1 one has considered in the literature (for example, Nataf 1948, or Gorman 1953) a restriction on the domain of individual income:

$$X(a, b) := \left\{ (x^h) \in \mathbb{R}_+^H \mid 0 < a \le x^h \le b \le \infty \right\}, a < b.$$

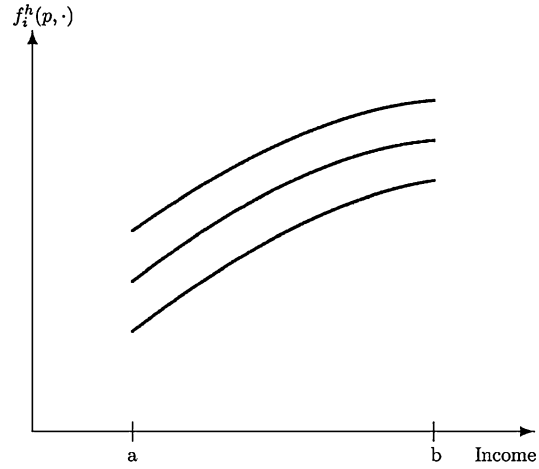
Proposition 2 shows that this restriction allows merely for some very limited and quite special heterogeneity in demand behaviour of the population H.

Proposition 2

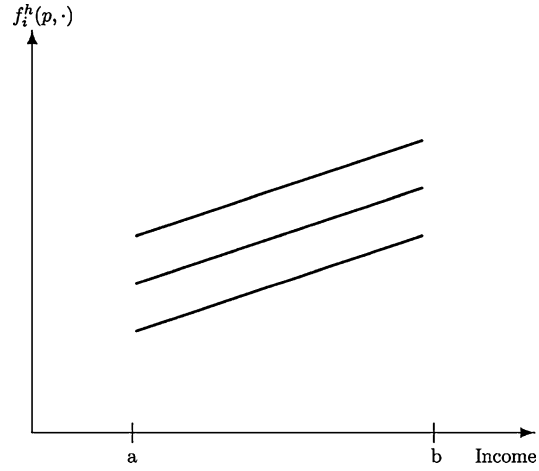
1. *There exists a function $F(p, G)$ such that (2) holds on $X(a, b) \times P$ if and only if for every commodity i and p A P the income expansion paths $f_i^h(p, \cdot), h \in H$, are parallel (vertically) on the interval (a,b); (with differentiability) $\partial_x f_i^h(p, x)$ does not depend on $h \in H$* (Fig. 1).

2. *There exists a function $F(p, \bar{x})$ such that (1) holds on $X(a, b) \times P$ if and only if for every commodity i and $p \in P$ the income expansion paths $f_i^h(p, \cdot)$, $h \in H$, are affine and parallel on the interval (a,b); (with differentiability) $\partial_x f_i^h(p, x)$ does not depend on $h \in H$ and $x \in (a,b)$* (Fig. 2).

3. *If all individual demand functions f^h belong to $\mathcal{F}$ and are homogeneous in (p,x), then the necessary condition in (i) implies that $f^h \equiv f$.*



Aggregation (Theory), Fig. 1



Aggregation (Theory), Fig. 2

Proof

(i) Consider any two households k and j in H and an income assignment in $X(a, b)$ with $x^k \neq x^j$. Now one interchanges the income of households k and j. This does not change the income distribution function. Hence, property (2) implies $f^k(p, x^k) + f^j(p, x^j) = f^k(p, x^j) + f^j(p, x^k)$. Thus $f^k(p, x^k) - f^k(p, x^j) = f^j(p, x^k) - f^j(p, x^j)$. Since it holds for all $x^k, x^j \in (a, b)$ and all p A P one obtains the claimed property in (i). The converse is trivial.

(ii) Instead of interchanging the income of households k and j one chooses $x^k + \Delta$ and $x^j + \Delta \in (a, b)$ for sufficiently small Δ. Property (1) then implies $f^k(p, x^k + \Delta) - f^k(p, x^k) = f^j(p, x^j) - f^j(p, x^j - \Delta) = f^k(p, x^j) - f^k(p, x^j - \Delta)$ by (i), which implies the claimed property in (ii). The converse is trivial.

(iii) If the expansion paths $f_i^k(p,\cdot)$, $h \in H$, are parallel on (a,b) for every $p \in P$, then homogeneity implies that they are also parallel on $(\lambda a,\lambda b)$ for all $\lambda > 0$ and $p \in P$. Hence they are parallel on $(0,\infty)$ for all $p \in P$. Continuity and $f^h(p, 0) = 0$ then implies the claim.

An alternative approach to allow for heterogeneous populations consists of considering, in addition to income, further explanatory variables for household demand. For example, in applications it is standard practice to stratify the whole population H by a certain profile $a = (a_1, a_2, \ldots)$ of observable household attributes, such as household size, age of household head, etc. Let $H(a)$ denote the sub-population of all households in H with attribute profile a. Without loss of generality one can assume that $a \in \mathbb{R}^m$. Let $G_{x,a}$ denote the joint distribution of function of x^h, a_h across H. Analogously to Proposition 1 one shows

Proposition 1 *There exists a function $F(p, G_x, a)$ such that*

$$\frac{1}{H}\sum_{h\in H} f^h(p,x^h) = F(p,G_{x,a})$$

for all conceivable income-attribute assignments and all $p \in P$ if and only if all sub-populations $H(a)$ are homogeneous in demand behaviour, that is, $f^h = f^a$ for all $h \in H(a)$.

Thus, the whole population need not be homogeneous, yet the joint distribution of x^h and a^h across H has typically a complex dependence structure, and hence, it cannot be modelled by some few parameters, as in the case of income.

Exact Income Aggregation

In the literature on 'exact income aggregation', as initiated by Gorman (1953), Lau (1982), and Jorgensen et al. (1982), one seeks for a representation of mean demand which is less restrictive than (3), yet more demanding than (4), that is to say $\frac{1}{H}\sum_{h\in H} f^h(p,x^h) = F(p, S_1(G_x), \ldots, S_N(G_x))$ on $\mathbb{R}_+^H \times P$ for some continuous function F from $P \times \mathbb{R}^N$ into $\mathbb{R}^l$ (the commodity space) and some vector of distributional statistics $S_1(G_x)$, ..., $S_N(G_x)$ with $N < H$. This representation is more demanding than (4); it does not require the knowledge of the entire income distribution since $N < H$.

If such a representation exists, then by Proposition 1, $f^h = f$, $h \in H$, and f is called 'exactly aggregable'. Thus, the question is whether there are exactly aggregable demand functions which are not linear in income and satisfy the basic restrictions of demand theory?

To simplify the presentation one assumes that all distributional statistics are 'generalized moments', that is, $S_n(G_x) = \int s_n(\xi) dG_x(\xi)$, with continuous functions $s_n(\,)$. Without loss of generality one can require that $s_n(0) = 0$.

Proposition 3 *There exists a representation of mean demand of the form*

$$\int f(p,\xi)dG_x(\xi) = F\left(p, \int s_1(\xi)dG_x(\xi), \ldots, \int s_N(\xi)dG_x(\xi)\right) \quad (5)$$

which holds for every income distribution function G_x of every finite population H and every price vector in P if and only if the function f is of the form

$$f(p,\xi) = \alpha_1(p)s_1(\xi) + \cdots \alpha_N(p)s_N(\xi), p \in P \text{ and } \xi \in \mathbb{R}_+ \tag{6}$$

where $\alpha_n(p) \in \mathbb{R}^l$.

Proof Trivially, (6) implies (5). Assume that (5) holds. Let $\mathcal{G}$ denote the set of all income distribution functions for every finite population. Note that for every G^1; $G^2 \in$ G and any rational λ with $0 \leq \lambda \leq 1$ it follows that $G^\lambda = \lambda G^1 + (1-\lambda)G^2 \in \mathcal{G}$. The representation (5) implies for every commodity i

$$f_i(p,\xi) = F_i(p, s_1(\xi), \ldots, s_N(\xi)), \quad p \in P \text{ and } \xi \in \mathbb{R}_+ \tag{7}$$

Now one shows that the function $F_i(p, \cdot)$ has a 'linear structure' on its relevant domain $\mathcal{D} := \{y \in \mathbb{R}^N | y_n = \int s_n(\xi)dG(\xi), G \in \mathcal{G},, n = 1, \ldots, N\}$, that is,

$$F_i\left(p, \lambda y^1 + (1-\lambda)y^2\right) = \lambda F_i\left(p, y^1\right) + (1-\lambda)F_i\left(p, y^2\right) \tag{8}$$

for every $y^1, y^2 \in$ D and any rational λ with $0 \leq \lambda \leq 1$. Indeed, $y_n^k = \int s_n(\xi)dG^k(\xi)$, k = 1,2 for some G^1, $G^2 \in$ G. Let $G^\lambda = \lambda G^{-1} + (1-\lambda)G^2$. Then $\int s_n(\xi)dG^\lambda(\xi) = \lambda \int s_n(\xi)dG^{-1}(\xi) + (1-\lambda)$. Hence $\lambda y^1 + (1-\lambda)y^2 \in \mathcal{D}$ since $G^\lambda \in$ G for rational λ. Consequently, the closure $\overline{\mathcal{D}}$ of $\mathcal{D}$ is convex. Since $G^\lambda \in$ G, one obtains from (5)

$$\int f_i(p,\xi)dG^\lambda(\xi) = F\left(p, \int s_1(\xi)dG^\lambda(\xi), \ldots, \int s_N(\xi)dG^\lambda(\xi)\right)_i = F_i(p, \lambda y^1 + (1-\lambda)y^2)$$

The left hand is equal to $\lambda \int f_i(p, \xi)dG^1(\xi) + (1-\lambda) \int f_i(p, \xi)dG^2(\xi) = \lambda F_i(p, y^1) + (1-\lambda)F_i(p, y^2)$ by (5), which proves (8). Since F_i is continuous, the 'linear structure' (8) also holds for any y^1, y^2 in the closure $\overline{\mathcal{D}}$ of $\mathcal{D}$ and any λ with $0 \leq \lambda \leq 1$. Since $s_n(0) = 0$ and $f(p, 0) = 0$ it follows from (7) that $F_i(p, 0) = 0$. Consequently, by (8), the restriction of the function $F_i(p, \cdot)$ on the convex domain $\overline{\mathcal{D}}$ can be extended to a function $\tilde{F}_I(p,\cdot)$, which is linear in y, that is, $\tilde{F}_i(p, y) = \alpha_1^i(p)y_1 + \cdots \alpha_N^i(p)y_N$. Thus (7) implies (6). The extension is unique if the dimension of the convex domain $\overline{\mathcal{D}}$ is equal to N.

Remark The proof of Proposition 3 is quite simple since it was assumed that the representation (5) must hold for all income distribution functions for *all* finite populations. This case is also treated in Heineke and Shefrin (1988), their proof, however, requires differentiability. If one only requires (5) to hold for all income distribution functions of a *given* population H with $N < H$, then it is much more difficult to obtain (6). See Lau (1982) and Heineke and Shefrin (1988).

Note that the global structural specification (6) is very restrictive if the demand function $f \in$ F has the basic properties of static demand theory. In fact, Heineke and Shefrin (1987) show the following result: *if $f \in$ F satisfies the budget-identity, is homogeneous in p and x and if no budget share function $w_i(p, x) : \equiv p_i f_i(p, x)/x$ is oscillating (that is, the derivative $\partial_x w_i(p, x)$ changes infinitely often its sign), then (6) implies $f(p, x) = \alpha(p)x$.*

Indeed, if $f \in$ F satisfies the budget identity, then $0 \leq w_i(p, x) \leq 1$. Let the budget share function $w_k(\overline{p},\cdot)$ be non-constant and non-oscillating. Consider the function $\varphi_\lambda(\cdot)$, $\lambda > 0$, defined by $\varphi_\lambda(x) = w_k(\overline{p}, \lambda x)$, and the linear function space which is generated by all functions $\varphi_\lambda(\cdot)$, $\lambda > 0$. Heineke and Shefrin (1987) argue that the dimension of this linear space is infinite. By homogeneity, $\varphi_\lambda(x) = w_k(\overline{p}/\lambda, x)$; thus, the linear space L which is generated by all budget share functions $w_k(p, \cdot)$, $p \in P$ has infinite dimension. Consequently, the demand function f cannot satisfy (6), since (6) implies that dim $\mathcal{L} \leq N$. Thus, if f satisfies (6) and $w_k(\overline{p},\cdot)$ is non-oscillating, then it must be constant, that is, $f_k(\overline{p},\cdot)$ is linear.

As a consequence, for demand functions which have the basic properties of atemporal demand

theory including non-oscillating budget share functions, one either has to be satisfied with a representation as in Proposition 1 or one is in the trivial case of Antonelli's Theorem.

Heterogeneous Populations

The representations (3), (4), and (5) of mean demand which have been considered up to now imply that the population of households must be homogeneous in demand behaviour, that is, $f^h \equiv f, h \in H$. The reason for this unsatisfactory fact is due to the very strong requirement that the representations must hold for every conceivable income assignment. This is more demanding than is needed in many applications, since there, changes in individual income are not entirely arbitrary; they might be the result of an underlying process. This point was emphasized by Malinvaud (1956) and (Malinvaud 1993). To capture this idea, one starts from an initial income assignment (x_0^h) (status quo), and then one considers a sequence $(x_n^h), n = 1, 2, \ldots$ or a set $X(x_0)$ of income assignments which are viewed as the result of the underlying (unspecified) process.

Which properties must the sequence (x_0^h) or the set $X(x_0)$ have such that for any assignment of demand functions f^h the representations of mean demand hold along this sequence or on the set $X(x_0)$?

We give three examples. The first one is well-known. The second and third example generalize substantially the first one.

Example 1 *Fixed income shares*

Starting from an initial income assignment (x_0^h), one defines the set $X(\delta) \subset \mathbb{R}_+^H$ of income assignments

$$X(\delta) := \left\{ (x^h) \in \mathbb{R}_+^H \middle| \, x^h/\overline{x} = x_0^h/\overline{x}_0 =: \delta^h \right\}$$

where $\overline{x}$ denotes mean income across H.

Given any assignment of demand functions f^h, $h \in H$, there exists a function F from $P \times \mathbb{R}_+$ into $\mathbb{R}_+^l$ such that mean demand has the representation

$$\frac{1}{H} \sum_{h \in H} f^h(p, x^h) \equiv F(p, \overline{x}) \text{ on } X(\delta) \times P \qquad (9)$$

The function F is defined by $F(p,\overline{x}) = \frac{1}{H} \sum_{h \in H} f^h(p, \delta^h \overline{x})$. If all f^h are linear in income then $F(p,\overline{x})$ is linear in mean income $\overline{x}$. Moreover, Eisenberg (1961) and Chipman and Moore (1979) have shown: if all f^h are generated by a utility function homogeneous of degree one then $F(p,\overline{x})$ is also generated by a utility function homogeneous of degree one given by

$$u(z) = \max_{z^h \in \mathbb{R}_+^l, \sum_H z^h = z} \prod_{h \in H} \left(u^h(z^h)\right)^{x_0^h/\overline{x}_0}$$

Example 2 *Rank preserving income changes*

Starting from an initial income assignment $(x_0^h)_{h \in H}$ one defines the set $X(x_0) \subset \mathbb{R}_+^H$ of income assignments (x^h) which have the property that every household keeps his rank position of income, that is, if for two households j and k, $x_0^j = x_0^k$ then $x^j = x^k$ and if $x_0^j < x_0^k$ then $x^i = x^k$ For any (x_1^h) and (x_2^h) in $X(x_0)$ there is a strictly increasing function ϕ such that $\phi(x_1^h) = x_2^h$, $h \in H$. Examples for $\phi(\ )$ are given in Fig. 3 (low income is increased, high income is decreased) and Fig. 4 (low and high incomes are decreased, middle ones increased) below.

Note that $(x^h) \in X(x_0)$ implies $X(x) = X(x_0)$ and $(x^h) \notin X(x_0)$ implies $X(x) \cap X(x_0) = \emptyset$. Thus, there is a finite partition $\{\tilde{X}_i\}$ of $\mathbb{R}_+^H$ into sets $\tilde{X}_i$ of rank preserving income assignments.

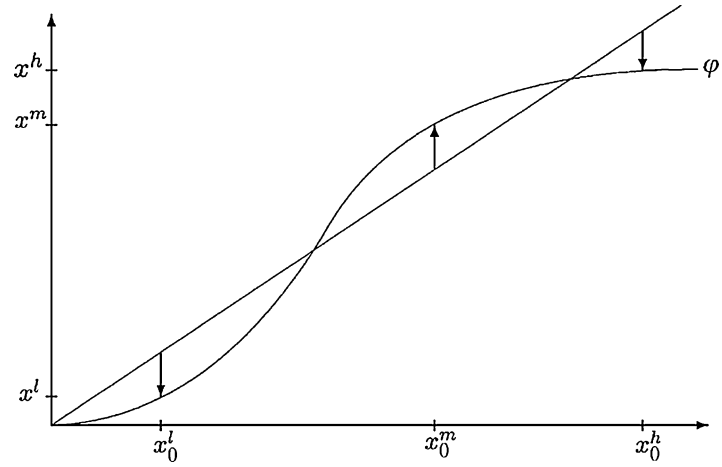
Note that for any rank preserving income assignments (x^h) in $X(x_0)$ one can recover the income assignment from knowing only its distribution function G_x, since $x^h = G_x^{-1} G_0(x_0^h)$ for every $h \in H$, where G^{-1} denotes the quantile function (quasiinverse) of the distribution function G, which is defined by.

$G^-(q) := \inf \{x \in \mathbb{R}_i | \, G(x) \geq q\}$. Consequently, one obtains:

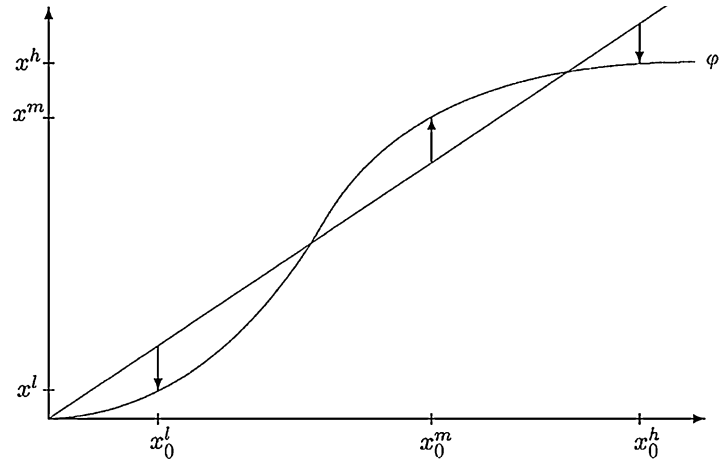
Given any assignment of demand functions f^h, $h \in H$, there exists a function $F(p,G)$ such that mean demand has the representation

$$\frac{1}{H} \sum_{h \in H} f^h(p, x^h) \equiv F(p, G_x) \text{ on } X(x_0) \times P. \quad (10)$$

Aggregation (Theory), Fig. 3



Aggregation (Theory), Fig. 4



The function F is defined by $F(p, G_x) = \frac{1}{H}\sum_{h \in H} f^h\left(p, G_x^{-1} G_0\left(x_0^h\right)\right)$.

There might be larger sets than $X(x_0)$ for which the representation (10) holds. For example, if households k and j have the same demand function then one can interchange their rank position. Thus, in defining a set X for which (10) holds, one should take into account the heterogeneity structure of $(f^h)_{h \in H}$. This is done in the next example

Example 3 *Common copula*

Let $\{f_1, \ldots, f_N\}$ be the set of distinct demand functions of the given assignment $(f^h)_{h \in H}$ Thus, for $h \in H$ there is an integer $n(h) \leq N$ such that $f^h = f_{n(h)}$. For every income assignment $(x^h)_{h \in H}$ consider the bivariate distribution function D_x, which is defined by

$$D_x(\xi, \eta) : = \frac{1}{H}\left\{h \in H \middle| x^h \leq \xi \text{ and } n(h) \leq \eta\right\}, \xi, \eta \in \mathbb{R}$$

The distribution function D_x and the price vector p determines mean demand $\frac{1}{H}\sum_{h \in H} f^h\left(p, x^h\right)$. The marginal distribution functions of D_x are denoted by G_x and V.

By Sklar's Theorem (see, for example, Nelson 1999), for every bivariate distribution function D with marginals G and V, there exists a copula

C (a function from $[0,1]^2$ into $[0,1]$ with certain properties) such that $D(\xi, \eta) = C(G(\xi), V(\eta))$ for all $\xi;\eta \in \mathbb{R}$. Conversely, if C is a copula and G and V are distribution functions, then $C(G(\xi), V(\eta))$ is a bivariate distribution function. Thus, a copula 'couples' the marginals to the bivariate distribution. The copula models the dependence structure of the bivariate distribution function.

Starting from an initial income assignment $\left(x_0^h\right)$, one considers the set $X(x_0,f) \subset \mathbb{R}_+^H$ of income assignments (x^h) such that the corresponding bivariate distribution functions D_x have a common copula. Thus, the dependence structure of (x^h, f^h) across H is the same for all (x^h) in $X(x_0, f)$. It follows that income assignments in the set $X(x_0)$ of rank preserving income assignments is contained in the set $X(x_0, f)$. Furthermore, *given any assignment of demand functions* $(f^h)_{h \in H}$, *there exists a function* $F(p, G)$ *such that mean demand has the representation*

$$\frac{1}{H}\sum_{h\in H} f^h\left(p,x^h\right) \equiv F(p,G_x) \text{ on} X(x_0,f) \times P$$

There is a very simple, however, special case which is worthwhile to be mentioned (and could have been discussed at the beginning). If the initial income x_0^h and the demand function f^h of household h are independently distributed across H, that is, $D_{x_0}(\xi, \eta) \equiv G_{x_0}(\xi)V(\eta)$ (the copula of D_{x_o} is equal to $C(u, v) = u \cdot v$), then the set $X(x_0,f) =: \mathcal{L}(x_0)$ is very large; it consists of all income assignments $\left(x^h\right) \in \mathbb{R}_+^H$ with the property: $x_0^k = x_0^j$ implies $x^k = x^j$. Then, one obtains

$$\frac{1}{H}\sum_{h\in H} f^h\left(p,x^h\right) \equiv F(p,G_x) \text{ on } \mathcal{L}(x_0) \times P$$

with $F(p,G) = \int \overline{f}(p,\xi)dG(\xi)$ where $\overline{f}(p,\xi) = \frac{1}{H}\sum_{h\in H} f^h\left(p,\overrightarrow{\xi}\right)$.

Monotone Mean Demand

The '*law*' *of demand* for a population of households asserts that the vector of price changes $\Delta F \in \mathbb{R}^l$ and the resulting vector of mean demand changes $\Delta F \in \mathbb{R}^l$ point in opposite directions, provided the price changes do not affect households' incomes (total expenditure) and demand functions (preferences). Thus, the 'law' asserts that the mean demand function $F(p)$ is *strictly monotone*, that is,

$$(p-q)\cdot(F(p)-F(q)) < 0 \text{ for all} p,q \in \mathbb{R}_{++}^l, p \neq q.$$

Strict monotonicity of mean demand implies, in particular, that for every commodity i the partial mean demand function F_i is strictly decreasing in its own price p_i and that the mean demand function $F(\ )$ is invertible (existence of an inverse demand function).

The goal of aggregation theory is to derive strict monotonicity of mean demand without assuming that households' demand functions $f^h(p, x)$ are strictly monotone in p.

Demand functions $f^h \in \mathcal{F}$ are assumed to be continuous in p and x and satisfy the budget-identity $p \cdot f(p, x) = x$. The function $f \in$ F satisfies the *Weak Axiom* of revealed preferences if for every price-income pair (p,x) and (p',x'), $p \cdot f(p',x') \leq x$ implies $p \cdot f(p, x) \geq x'$, and satisfies the *Axiom* of revealed preferences, if $f(p, x) \neq f(p', x')$ and $p \cdot f(p', x') \leq x$ implies $p' \cdot f(p, x) > x'$.

Every demand function which is derived from a continuous, strictly convex and non-saturated preference relation satisfies the Axiom, yet it is not necessarily monotone.

Theorem (Hildenbrand 1983)

1. The function $F(p) := \int_0^\infty f(p,x)\rho(x)dx$ is monotone, that is, $(p-q)\cdot(F(p)-F(q)) \leq 0$ for all p, q in $\mathbb{R}_{++}^l$, if $f \in$ F satisfies the Weak Axiom of revealed preferences and ρ is a density which is non-increasing on $\mathbb{R}_+$ with $\int_0^\infty \rho(x)dx < \infty$.
2. The function F is strictly monotone, if, in addition, f satisfies the Axiom of revealed preferences and the expansion paths $f(p, .)$ and $f(q, .)$ have only 0 in common for any p, q that are not collinear.

Interpretation The underlying micro-model is a population H of households which is 'indefinitely large'; mathematically, an atomless measure space, for example, the unit interval [0,1] with Lebesgue measure. Every household $h \in [0,1]$ is modelled by its income $x(h) \geq 0$ and the common demand function f. The income assignment $x(\,)$ is an integrable function whose distribution admits a density ρ. Thus, mean demand $F(p) = \int_0^1 f(p, x(h))dh = \int_0^\infty f(p,x)\rho(x)dx$.

Three qsts are relevant:

1. Why a continuum of households? Does the result still hold approximately for a large but finite population?
2. Why a non-increasing income density? Does monotonicity of F fail if the density is first increasing and then decreasing?
3. Why a common demand function? Does the result extend to heterogeneous populations in income and demand behaviour?

The discussion of these qsts is simplified by assuming that f is continuously differentiable in p and x. Then monotonicity of F is equivalent with negative semi-definiteness (n.s.d.) of the Jacobian matrix $\partial_p F(p)$ for all p, that is, $\sum_{i,j=1}^{l} v_i v_j \, \partial_{p_i} F_j(p) \leq 0$ for all $v \in \mathbb{R}^l$, and the Weak Axiom for f is equivalent with n.s.d. of the Slutzky substitution matrix. Consequently, monotonicity of F follows from the positive semi-definiteness (p.s.d.) of the mean income effect matrix $I(f, \rho) = \int I(f, x)\rho(x)dx$, where $I(f, x) = (f_i(p, x)\partial_x f_j(p, x))_{i,j=1,\ldots,l}$.

Question 1 The mean income effect matrix for a finite population H, that is, $\frac{1}{H}\sum_H \left(f_i(p,x^h)\cdot \partial_x f_j(p,x^h)\right)_{i,j} = I_H$ is p.s.d. if and only if for every $v \in \mathbb{R}^l$, $v \cdot I_H v = \frac{1}{H}\sum_H g'(x^h) \geq 0$ where $g(x) := \frac{1}{2}(v \cdot f(p,x))^2$. Assume that income x^h is measured in multiples of Δ (euro). Let $\pi_n := \frac{1}{H}\left\{h \in H \middle| x^h = n \cdot \Delta =: x_n\right\}, n = 0, 1, \ldots$ Then

(1) $\frac{1}{H}\sum_H g'(x^h) = \sum_{n=0}^{\infty} \pi_n g'(x_n) = \sum_{n=1}^{\infty} \frac{1}{\Delta}(\pi_{n-1} - \pi_n)g(x_n) + o(\Delta)$ using the approximation

(2) $g'(x_n) = \frac{1}{\Delta}(g(x_{n+1}) - g(x_n)) + o(\Delta)$.

Consequently, one needs $\pi_{n-1} \geq \pi_n$, $n = 1,\ldots$, to obtain a non-negative first term on the right hand side of (1); this is the finite analogue of a non-increasing density. *Thus, for a finite population with a small* Δ (which requires by $\pi_{n-1} \geq \pi_n$ a large population) *one obtains the desired result up to the small term* $o(\Delta)$. For a population $H = [0,1]$ one does not need the approximation (2) and hence $o(\Delta)$, since (1) becomes $\int g'(x)\rho(x)dx = -\int g(x)\rho'(x)dx$ (by partial integration), which is non-negative for a non-increasing differentiable density ρ.

Question 2 The mean income effect matrix $I(f, \rho)$ is p.s.d. in each of the two extreme cases: either, ρ is non-increasing and no assumption on the shape of the income expansion path $f_i(p, \cdot)$ or, no assumption on ρ yet linearity of $f_i(p, \cdot)$. There must be results in between. Indeed, if the curvature of all income expansion paths $f_i(p, \cdot)$ is limited and the unimodal density ρ is sufficiently skewed, then $I(f, \rho)$ is p.s.d.

Example All income expansion paths restricted to the interval $[0\bar{x}]$ are polynomials of degree n (note that, no non-linear $f_i(p, \cdot)$ can be a polynomial on $\mathbb{R}_+$) and ρ is concentrated on $[0\bar{x}]$. Then, $I(f,g)$ is p.s.d. if and only if the matrix $M(n, \rho) := ((i+j)m_{i+j-1})_{i,j=1,\ldots,n}$ is p.s.d. where $m_k := \int x^k \rho(x)dx$ (Hildenbrand 1994, Appendix 6).

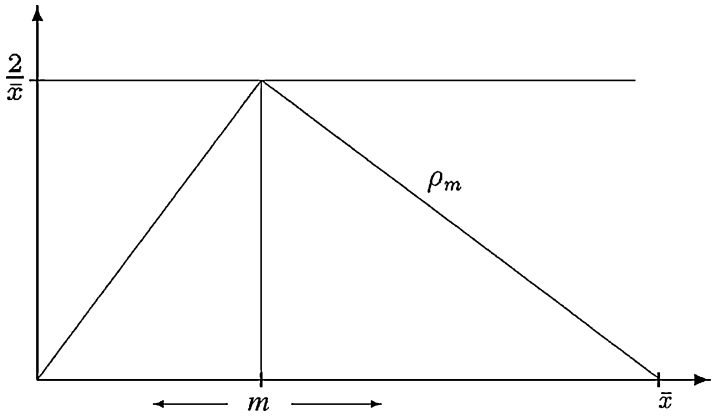
Let the densities ρ_m be as in Fig. 5.

For every n there exists $m(n) > 0$ *such that* $I(f, \rho_m)$ *is p.s.d. if* $m \leq m(n)$; for example, $n = 2$, $m(2) = 0.38\bar{x}$ or $n = 3, m(3) = 0.14\bar{x}$.

For a more general analysis see Chiappori (1985) and Hildenbrand (1994).

Question 3 A population of households that is *heterogeneous in income and demand* functions is described by a joint distribution μ of income and demand functions, that is, μ is a distribution on $\mathbb{R}_+ \times \mathcal{F}$. (A reader not familiar with distributions on function spaces might replace $\mathcal{F}$ by a finite set $\mathcal{F}_0$.) As before, the marginal distribution of income admits a density ρ. The conditional distribution of demand functions given the income level x is denoted by $\nu(x)$. Then *mean demand*

Aggregation (Theory), Fig. 5



$$F(p) := \int_{\mathbb{R}_+ \times \mathcal{F}} f(p,x)d\mu = \int_0^\infty \overline{f}(p,x)\rho(x)dx$$

where $\overline{f}(p,x) := \int_{\mathcal{F}} f(p,x)dv(x)$. Consequently, the Theorem or the extensions discussed under Question 2 imply that $F(p)$ is monotone provided the function $\overline{f}$ satisfies the Weak Axiom. This approach to derive monotonicity for a heterogeneous population is the most direct, yet not the most general way (see Hildenbrand 1994).

It is well-known (Hicks 1956, p. 53) that $\overline{f}$ does not necessarily satisfy the Weak Axiom, even if individual demand functions are derived from utility maximization. The following two assumptions (which, again, are not the most general ones) imply that $\overline{f}$ satisfies the Weak Axiom

(a) *Independence*: $v(x)$ does not depend on x
(b) *Increasing dispersion*: the distribution $D(x + \Delta)$; $\Delta > 0$, is more dispersed than the distribution $D(x)$, where $D(\xi)$ denotes the distribution (in the commodity space $\mathbb{R}^l$) of individual demand of all households with income ξ at the price p (that is, $D(\xi)$ is the image distribution of v under the mapping $f \mapsto f(p, \xi)$).

Generalizing the one-dimensional case where the variance is a measure of dispersion one chooses the positive definiteness of the covariance matrix as a measure of dispersion for distributions on $\mathbb{R}^l$. Thus, increasing dispersion means that for $\Delta > 0$, $covD(x + \Delta) - covD(x)$ is positive semi-definite.

Assumptions (a) and (b) are quite restrictive, in particular, the independence assumption. Therefore one partitions the whole population H into sub-populations $H(a)$ by stratifying with respect to a certain vector a of household attributes (household size, age,...) and then one requires assumptions (a) and (b) for each sub-population $H(a)$. The role of stratifying is to reduce the heterogeneity in demand behaviour. In the extreme case, where stratifying leads to a homogeneous sub-population in demand behaviour, assumptions (a) and (b) are trivially satisfied. If the income density of each sub-population $H(a)$ is non-increasing on R_b or if the extension discussed in Question 2 apply, the mean demand of each sub-population is monotone and hence also the mean demand of the whole population, since monotonicity is additive.

A more general definition of 'increasing dispersion' and a detailed discussion is given in Hildenbrand (1994). For an empirical study of the law of demand, see Härdle et al. (1991).

A broader discussion of the law of demand and related properties including cases where income is price dependent is contained in the entry law of demand.

See Also

- ▶ Aggregation (Econometrics)
- ▶ Copulas
- ▶ Law of Demand

Bibliography

Antonelli, G.B. 1886. *Sulla teoria matematica della economia politica*. Pisa: Nella Tipografia del Folchetto. Trans. J.S. Chipman and A.P. Kirman in *Preferences, utility and demand*, ed. J.S. Chipman, L. Hurwicz and M.K. Richter. New York: Harcourt Brace Jovanovich, 1971.

Chiappori, P. 1985. Distribution of income and the law of demand. *Econometrica* 53: 109–127.

Chipman, J.S., and J. Moore. 1979. On social welfare functions and the aggregation of preferences. *Journal of Economic Theory* 21: 111–139.

Eisenberg, B. 1961. Aggregation of utility functions. *Management Science* 7: 337–350.

Gorman, W.M. 1953. Community preference fields. *Econometrica* 21: 63–80.

Grandmont, J.M. 1992. Transformations of the commodity space, behavioural heterogeneity, and the aggregation problem. *Journal of Economic Theory* 57: 1–35.

Härdle, W., W. Hildenbrand, and M. Jerison. 1991. Empirical evidence on the law of demand. *Econometrica* 59: 1525–1549.

Heineke, J., and H. Shefrin. 1987. On some global properties of Gorman class demand systems. *Economics Letters* 25: 155–160.

Heineke, J., and H. Shefrin. 1988. Exact aggregation and the finite basis property. *International Economic Review* 29: 525–538.

Hicks, J.R. 1956. *A revision of demand theory.* London: Oxford University Press.

Hildenbrand, W. 1983. On the law of demand. *Econometrica* 51: 997–1019.

Hildenbrand, W. 1994. *Market demand*. Princeton: Princeton University Press.

Hildenbrand, W., and A. Kneip. 2005. On behavioral heterogeneity. *Economic Theory* 25: 155–169.

Jorgensen, D.W., L. Lau, and T.M. Stoker. 1982. The transcendental logarithmic model of aggregate consumer behavior. In *Advances in econometrics*, ed. R.L. Basmann and G.F. Rhodes. Greenwich, CT: JAI Press.

Lau, L. 1982. A note on the fundamental theorem of exact aggregation. *Economics Letters* 9: 119–126.

Malinvaud, E. 1956. L'agregation dans les modèles économiques. *Cahiers du Seminaire d'Econométrie* 4: 69–146.

Malinvaud, E. 1993. A framework for aggregation theories. *Ricerche Economiche* 47(2): 107–135.

Mitjuschin, L.G., and W.M. Polterovich. 1978. Criteria for monotonicity of demand functions [in Russian]. *Ekonomika i Matematicheskie Metody* 14: 122–128.

Nataf, A. 1948. Sur la possibilité de construction de certains macromodèles. *Econometrica* 16: 232–244.

Nelson, R. 1999. An introduction to copulas. In *Lecture notes in statistics 139*. New York: Springer Verlag.

Trockel, W. 1984. Market demand: An analysis of large economies with nonconvex preferences. In *Lecture notes in economics and mathematical systems 223*. Heidelberg: Springer Verlag.

Aggregation of Economic Relations

Walter D. Fisher

A simplification or aggregation problem is faced by a research worker whenever he finds that his data are too numerous or in too much detail to be manageable and feels the need to reduce or combine the detailed data in some manner. He will want to use aggregative measures for his new groups. But what will be the effect of this procedure on his results? How can he choose among alternative procedures? In grouping his data and/or relations he must also decide how many groups to use; a smaller number is more manageable but will cause more of the original information to be lost. The research worker seeks a solution of this problem that will best serve his objectives, or those of some decision-marker who will use his results.

For example, say that a true micro-model is

$$y = Px + v \tag{1}$$

where y is vector of g endogenous variables and x is a vector of h predetermined variables. It is desired to work with a macro-model

$$\overline{y} = \overline{P}\overline{x} + \overline{v} \tag{2}$$

where $\overline{y}$ is a reduced *vector* of f aggregated endogenous variables, and $\overline{x}$ is a reduced vector of j predetermined variables, where $f < g$, and $j < h$. The reduction is to be made in such a manner that when predictions are made with the macro-model, results will be as close as possible to those that could have been obtained with the micro-model.

General reviews of the aggregation problem in the various stages of its development may be found in Malinvaud (1956), Theil (1962), Fisher (1969) and Chipman (1976).

It is not surprising that the aggregation problem in economics began to attract attention with the development of econometrics since the task of inferring realistic models becomes particularly

acute when only limited empirical data are available. It was only a dozen years or so after the founding of the Econometric Society that there occurred an early methodological discussion of the problem in Klein (1946), May (1946, 1947) and Shou Shan Pu (1946). This discussion related specifically to simultaneous equation macro-models.

The important and pioneering work of Theil (1954) treated the question of the consequences of aggregation in a stochastic model of simultaneous equation. Theil derived relationships between estimated parameters in detailed and aggregated models when the parameters are estimated by linear unbiased methods.

The approach of Theil, that of measuring the consequences of aggregation in terms of the discrepancies from a true micro-model, leads directly to the goal of optimal aggregation – that is, selecting a mode of aggregation resulting in a macro-model based on the data at hand and a given degree of detail, such that the expected discrepancies are minimized. To pursue this goal it is necessary to postulate a loss function in terms of the discrepancies and to have available a procedure for minimizing expected loss from a very large number of alternatives. Say that it is desired to predict y in (1) with small error, but that it is also desired to use a simplified model, $\tilde{P}$, of the same size as P, so the prediction will be

$$\tilde{y} = \tilde{P}x. \tag{3}$$

The simplified model $\tilde{P}$ is considered to be subject to certain a priori restrictions. For example, it may be assumed to be of a rank lower than that of P, or to be expressible in the form

$$\tilde{P} = T'\overline{P} \tag{4}$$

where $\overline{P}$ is an *aggregated matrix* of smaller order than P, and T' is given a priori. Say that the cost of this procedure to the investigator is

$$c = E(\tilde{y} - y)'C(\tilde{y} - y), \tag{5}$$

where C is a known positive-definite matrix that weights the relative importance of forecast errors in the various endogenous variables and their interactions. It has been shown that

$$c = \mathrm{tr}\, C(\tilde{P} - P)M(\tilde{P} - P)' + \text{constant} \tag{6}$$

where $M = E(xx')$. The problem of choosing $\tilde{P}$ so as to minimize c may be called a *simplification problem*.

The lower the rank of $\tilde{P}$, or the smaller the dimensions of $\overline{P}$, the more severe is the aggregation and simplification. To find the matrix $\overline{P}$, or $\tilde{P}$ that minimizes the cost c subject to a given level of severity of aggregation, is a well defined but not a trivial problem. It may be accomplished in two steps: first, finding the optimal $\overline{P}$ conditional on a partition and second, searching for the partition that gives a minimum minimorum c. For the second step a computer is necessary. First suggested by Hurwicz (1952) and Malinvaud (1956), the optimal aggregation approach has been extended and applied to a number of econometric problems by Fisher (1953, 1962, 1969) and Chipman (1975a, b, 1976).

One of the most frequent applications and most strongly felt needs of aggregation is to Leontief inter-industry (input–output) models. We can make our equation (1) above into an input–output model by setting g = h, defining y as the set of outputs, x the set of final demands, and defining $P = (I - A)^{-1}$ where A is the matrix of technical coefficients. Here it is natural to require that the aggregation over both rows and columns of the matrix A involve the same partition, that is, that the combination of 'small industries' into 'large industries' implied by the row partition be the same as for the column partition. Some excellent preliminary discussion of the model is given by Leontief (1947). Conditions for obtaining perfect (without error) aggregation in this system were given by Hatanaka (1952).

Since the input–output model may be considered a special case of the simultaneous equations model, the same principles of optimal aggregation may be applied to find an aggregated or a simplified model. This approach is used in McCarthy (1956), Fisher (1958, 1969) and Neudecker (1970).

There is a well known correspondence between such concepts as distance, variance, and scatter, on the one hand, and entropy and information content on the other. If an m by n rectangular table contains a set X of numbers that sum to unity, the *entropy* of the table may be defined as

$$E(X) = -\sum_{i,\,j=1}^{n} x_{ij} \log x_{ij}. \tag{7}$$

This may be considered a measure of the degree of sameness or homogeneity of the elements of the matrix X.

If X is aggregated by rows and by columns, an aggregated entropy may be found from the aggregated cells of the smaller matrix. This entropy will be larger than that of X. The difference may be regarded as a loss of information from the aggregation. The problem may be posed: to find the mode of aggregation (to a specified degree of detail) that minimizes this loss.

Skolka (1964) and Theil (1967) have applied this idea to input–output tables. Fisher (1969, ch. 6) has shown an exact correspondence between this problem and the minimization of his objective function s. Recent insights into the aggregation problem in input–output analysis are found in Tintner and Sondermann (1977) and Laisney (1984).

Practically all of the work reviewed so far has proceeded on the assumption that the micromodel is true, or at least that the microdata with which the investigator works form an unbiased estimate of the truth. Thus, the expected loss from using an aggregated artefact can never be negative, and can be tolerated only if there is a compensating gain from aggregation, owing to increased manageability, understanding, etc., of a smaller model.

But in Grunfeld and Griliches (1960) an example was presented where the errors were *less* after aggregation. The monograph of Ringwald (1980) made the point that this situation is probably very frequent in economics, especially as so-called microdata have in reality undergone much processing and are a pre-aggregation of unobserved, yet more detailed data, probably subject to bias. Ringwald's critique has been followed up by Chipman (1985), who has developed formulae expressing the relationship between stage 1 and stage 2 models, where stage 1 is the result of some previous aggregation. The issue is obviously of considerable importance and it is evident that more work needs to be done.

See Also

▶ Separability

Bibliography

Chipman, J.S. 1975a. The aggregation problem in econometrics. *Advances in Applied Probability* 7: 72.

Chipman, J.S. 1975b. Optimal aggregation in large-scale economic models. *Sankhya* 37(4): 121–159.

Chipman, J.S. 1976. Estimation and aggregation in economics: An application of the theory of generalized inverses. In *Generalized inverses and applications*, ed. M.Z. Nashed. New York: Academic.

Chipman, J.S. 1985. Testing for reduction of mean-square error by aggregation in dynamic econometric models. In *Multivariate analysis VI: Proceedings of the sixth international symposium on multivariate analysis*. Amsterdam: North-Holland.

Day, R.H. 1963. On aggregating linear programming models of production. *Journal of Farm Economics* 45: 797–813.

Fei, J.C.H. 1956. A fundamental theorem for the aggregation problem of input–output analysis. *Econometrica* 24(4): 400–412.

Fisher, W.D. 1953. On a pooling problem from statistical decision viewpoint. *Econometrica* 21(4): 567–585.

Fisher, W.D. 1958. Criteria for aggregation in input–output analysis. *Review of Economics and Statistics* 40(3): 250–260.

Fisher, W.D. 1962. Optimal aggregation in multi-equation prediction models. *Econometrica* 30(4): 744–769.

Fisher, W.D. 1969. *Clustering and aggregation in economics*. Baltimore: The Johns Hopkins Press.

Fisher, W.D. 1979. A note on aggregation and disaggregation. *Econometrica* 47(3): 739–746.

Fisher, W.D., and P.L. Kelley. 1968. Selecting representative firms in linear programming. Ch. 13. In *Economic analysis and agricultural policy*, ed. R.H. Day. Ames: Iowa State University Press, 1982.

Gorman, W.M. 1968. The structure of utility functions. *Review of Economic Studies* 35: 367–390.

Grunfeld, Y., and Z. Griliches. 1960. Is aggregation necessarily bad? *Review of Economics and Statistics* 42(1): 1–13.

Hatanaka, M. 1952. Note on consolidation within a Leontief system. *Econometrica* 20(3): 301–303.
Hurwicz, L. 1952. Aggregation in macroeconomic models [abstract]. *Econometrica* 20(3): 489–491.
Ijiri, Y. 1971. Fundamental queries in aggregation theory. *Journal of the American Statistical Association* 66: 766–782.
Klein, L.R. 1946. Remarks on the theory of aggregation. *Econometrica* 14(4): 303–312.
Laisney, F. 1984. Theory and practice in optimal aggregation of linear models. *Economic Letters* 15: 315–324.
Leontief, W. 1947. Introduction to a theory of the internal structure of functional relationships. *Econometrica* 15(4): 361–373.
Malinvaud, E. 1956. *L'agrégation dans les modéles économiques*, Cahiers du Seminaire d'Econométrie, vol. 4. Paris: Centre National de la Recherche Scientifique.
May, K. 1946. The aggregation problem for a one-industry model. *Econometrica* 14(4): 285–298.
May, K. 1947. Technological change and aggregation. *Econometrica* 15(1): 51–63.
McCarthy, J. 1956. Aggregation in the Leontief model. Joint Allied Social Science Association meeting in Cleveland, Ohio, 27 December.
Nataf, A. 1960. Résultats et directions de recherche dans la théorie de l'agrégation. In *Logic, methodology and philosophy of science. Proceedings of the 1960 International Congress*, ed. E. Nagel, P. Suppes, and A. Tarski. Stanford: Stanford University Press, 1962.
Neudecker, H. 1970. Aggregation in input–output analysis: An extension of Fisher's method. *Econometrica* 38(6): 921–926.
Ringwald, K. 1980. *A critique of models in linear aggregation structures*. Boston: Oelgeschlager, Gunn & Hain.
Schneeweiss, H. 1965. Das Aggregationsproblem. *Statistische Hefte*. doi:10.1007/BF02922284.
Shou Shan Pu. 1946. A note on macroeconomics. *Econometrica* 14(4): 299–302.
Skolka, J. 1964. *The aggregation problem in input–output analysis*. Prague: Czechoslovakian Academy of Sciences.
Sondermann, D. 1973. Optimale Aggregation von grossen linearen Gleichungssystemen. *Zeitschrift für Nationalökonomie* 33(3–4): 235–250.
Theil, H. 1954. *Linear aggregation of economic relations*. Amsterdam: North-Holland.
Theil, H. 1962. Alternative approaches to the aggregation problem. In *Logic, methodology and philosophy of science. Proceedings of the 1960 International Congress*, ed. E. Nagel, P. Suppes, and A. Tarski. Stanford: Stanford University Press.
Theil, H. 1967. *Economics and information theory.* Chicago: Rand McNally.
Tintner, G., and D. Sondermann. 1977. Statistical aspects of economic aggregation. In *Mathematical economics and game theory*, ed. R. Henn and O. Moeschlin. Berlin: Springer.

Aggregation Problem

Franklin M. Fisher

Microeconomic theory elegantly treats the behaviour of optimizing individual agents in a world with an arbitrarily long list of individual commodities and prices. However, the desire to analyse the great aggregates of macro-economics-gross national product, inflation, unemployment, and so forth-leads to theories that treat such aggregates directly. What is the relation of such theory (or empirical work) to the underlying theory of the individual agent. When is it possible to speak of 'food' rather than of 'apples, bananas, carrots, etc.?' When can one treat the investment decisions of all firms together as though there were a single good called 'capital' and all firms were a single firm?

Leontief's Theorem

Underlying many results on aggregation is a theorem of Leontief (1947a, b). Let x and y be vectors of variables and $F(x,y)$ a twice-differentiable function. It is desired to aggregate over x, that is to replace x with a scalar aggregator function, $g(x)$, such that $F(x) \equiv H[g(x),y]$. This can be done if and only if, along any surface on which $F(x,y)$ is constant, the marginal rate of substitution between each pair of elements of x is independent of y.

Hicks-Leotief Aggregation

Since optimizing, price-taking agents equate marginal rates of substitution to price ratios, one restriction permitting aggregation over commodities is the assumption that the prices of all goods to be included in an aggregate always vary proportionally. This is called 'Hicks-Leontief aggregation' (Leontief 1936; Hicks 1939) and is a powerful expository tool. It requires no special assumptions as to the separability of utility or

production functions, but is only applicable in relatively artificial situations. Under more general circumstances, and especially where aggregation over agents is involved, restrictions on utility or production functions become essential.

Consumption

Consider a single household. Suppose that we wish to describe behaviour in terms of aggregate commodities such as 'food' or 'clothing'. By Leontief's Theorem, a food aggregate exists if and only if the marginal rate of substitution between any two kinds of food is independent of consumption of any non-food commodity. If a similar restrictive condition is satisfied for all the aggregates to be constructed, then the household's utility function can be written in aggregate terms.

Even such restrictive conditions will not always suffice. If we wish to represent the household as maximizing the aggregated utility function subject to an aggregated budget constraint, we must have aggregate prices as well as aggregate consumption goods. This requires that aggregates such as 'food' be homothetic in their component variables, again considerably restricting the household's utility function (Gorman 1959; Blackorby et al. 1970).

Aggregation over agents presents a different set of questions. Suppose that we wish to treat the aggregate demands of a collection of households as the demands of a single, aggregate household. Then only aggregate income and not its distribution can influence demand. At given prices, this makes the incomederivative of every household's demand for a given commodity the same constant. Engel curves must be parallel straight lines. If zero income implies zero consumption, then all households must have the same homothetic utility function (Gorman 1953).

In general, the only consumer-theoretic restrictions obeyed by aggregate demand functions are those of continuity, homogeneity of degree zero, and the various restrictions implied by the budget constraint. This corresponds to an important result (Sonnenschein 1972, 1973) of general equilibrium theory on aggregate excess demand functions with Walras' Law replacing the budget constraint.

Production

Aggregation over inputs or outputs for a single firm also requires restrictive Leontief conditions on marginal rates of substitution. Aggregation over firms, however, leads to richer results.

Assume that every firm produces the same output from the same two inputs, capital and labour. Define Y, K, and L, as the totals over firms of output, capital and labour, respectively. We wish to represent the aggregate technology of the entire economy as $Y = F(K,L)$.

At first glance, this seems to lead to the same sort of result as in the case of households. Transfers of labour among firms must leave total output unchanged. Hence each firm's production function must be linear in labour with all firms having the same coefficient. If a similar condition applies to capital, every firm must have the same linear production function. Apparently, aggregation is not generally possible even if all firms are the same! (Nataf 1948).

This formulation overlooks the fact that production functions involve efficiency conditions, giving *maximum* output for given inputs. If the total capital, K, and labour, L, available to the economy are assigned to firms to maximize total output, Y, then aggregate production can *always* be written as $Y = F(K,L)$. No further conditions are required.

This optimal-assignment is relatively natural for labour. It holds in competitive labour markets if all firms face the same wage (or in efficiently managed, centrally planned economies). It is not natural for capital, however, once we drop the assumption that all firms use the same type of physical capital. Suppose that technology is embodied in the capital stock and that capital cannot be shifted among firms. Then labour aggregation in this simple model remains easy, but capital aggregation is another matter.

Assume constant returns. If all firms had the same production function, differing amounts of capital would lead to differing amounts of labour

with the labour-capital ratio the same in all firms. Then firms would differ only as to scale, and constant returns would make many small firms equivalent to one large one, permitting aggregation.

Unfortunately, capital augmentation is the *only* case permitting capital aggregation under constant returns in this model. Extensions to allow more types of capital in a given firm lead to similar results, as well as requiring that individual production functions permit capital aggregation. The requirements for the existence of partial capital aggregates such as 'plant' and 'equipment' are also very restrictive (Fisher 1965, 1983; Gorman 1968).

Now suppose that firms do not all have the same production function but that (for every υ), the υth firm's production function can be written as $F(b_\upsilon K_\upsilon, L_\upsilon)$, where the function, $F(\cdot,\cdot)$ is common to all firms, but the parameter b_υ can differ. This 'capital-augmenting' case is very restrictive, making one unit of one type of capital the exact duplicate of a fixed number of units of another. Having a different type of capital is equivalent to having more of the same type, and the argument given above shows that aggregation is permitted. The aggregate production function is $F(J, L)$, where J is the sum of the terms $b_\upsilon k_\upsilon$ (Solow 1964).

Continue to assume capital to be firm-specific, but let there be several types of labour or of output. Output aggregation requires first that each firm's technology be separable in terms of output-the marginal rate of substitution between any pair of outputs must be independent of inputs. Further, under constant returns, the output-aggregator function must be the same for all firms (in contrast to the case of capital where production functions must be the same *after* capital aggregation). This means that firms cannot specialize; every firm must produce the same market-basket of outputs differing only as to scale. Similar conditions apply to labour (Fisher 1968).

Perhaps surprisingly, the restrictive nature of such results does not really depend on the assumption that capital is firm-specific, once we leave the expository case of one output, one kind of capital, and one labour input for each firm. In general, aggregation over any set of inputs or outputs requires separability in each firm's production function. Further, under constant returns, even if capital is not firm-specific, aggregation over firms requires either that the aggregator functions applied to the firms all be the same (no specialization) or, if not, that the *only* difference in production functions be the nature of the aggregator function (generalized capital augmentation) (Fisher 1982).

Abandoning constant returns does not provide practical help. Most non-constant returns cases do not permit aggregation even if all firms have the *same* production function. The cases that do are very restrictive (Fisher 1965, 1968; Gorman 1968; Blackorby and Schworm 1984).

Such results show that the analytic use of such aggregates as 'capital', 'output', 'labour' or 'investment' as though the production side of the economy could be treated as a single firm is without sound foundation. This has not discouraged macroeconomists from continuing to work in such terms.

See Also

► Separability

Bibliography

Blackorby, C., and W. Schworm. 1984. The structure of economies with aggregate measures of capital: A complete characterization. *Review of Economic Studies* 51(4): 633–650.

Blackorby, C., G. Lady, D. Nissen, and R.R. Russell. 1970. Homothetic separability and consumer budgeting. *Econometrica* 38(3): 468–472.

Fisher, F.M. 1965. Embodied technical change and the existence of an aggregate capital stock. *Review of Economic Studies* 32(4): 263–288.

Fisher, F.M. 1968. Embodied technology and the existence of labour and output aggregates. *Review of Economic Studies* 35(4): 391–412.

Fisher, F.M. 1982. Aggregate production functions revisited: The mobility of capital and the rigidity of thought. *Review of Economic Studies* 49(4): 615–626.

Fisher, F.M. 1983. On the simultaneous existence of full and partial capital aggregates. *Review of Economic Studies* 50(1): 197–208.

Gorman, W.M. 1953. Community preference fields. *Econometrica* 21(1): 63–80.
Gorman, W.M. 1959. Separable utility and aggregation. *Econometrica* 27(3): 469–481.
Gorman, W.M. 1968. Measuring the quantities of fixed factors. In *Value, capital, and growth: Papers in honour of Sir John Hicks*, ed. J.N. Wolfe. Edinburgh: University of Edinburgh Press.
Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon.
Leontief, W.W. 1936. Composite commodities and the problem of index numbers. *Econometrica* 4(1): 39–59.
Leontief, W.W. 1947a. A note on the interrelation of subsets of independent variables of a continuous function with continuous first derivatives. *Bulletin of the American Mathematical Society* 53: 343–356.
Leontief, W.W. 1947b. Introduction to a theory of the internal structure of functional relationships. *Econometrica* 15(4): 361–373.
Nataf, A. 1948. Sur la possibilité de construction de certains macromodèles. *Econometrica* 16(3): 232–244.
Solow, R.M. 1964. Capital, labor, and income in manufacturing. In *The behavior of income shares*, Studies in Income and Wealth, ed. R.M. Solow, 101–128. Princeton: Princeton University Press.
Sonnenschein, H. 1972. Market excess demand functions. *Econometrica* 40(3): 549–563.
Sonnenschein, H. 1973. Do Walras' identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6(4): 345–354.

Agricultural Economics

C. Ford Runge

Abstract

Agricultural economics arose in the late 19th century, combined the theory of the firm with marketing and organization theory, and developed throughout the 20th century largely as an empirical branch of general economics. The discipline was closely linked to empirical applications of mathematical statistics and made early and significant contributions to econometric methods. From the 1960s, as agricultural sectors in the OECD countries contracted, agricultural economists were drawn to the development problems of poor countries, to the trade and macroeconomic policy implications of agriculture in richer countries, and to a variety of issues in production, consumption, environmental and resource economics.

Keywords

Agricultural economics; Agricultural subsidies; Agriculture and economic development; American Agricultural Economics Association; American Economic Association; American Farm Economics Association; Australian Agricultural Economics Society; Bellman's dynamic programming principle; Black, J. D.; Boulding, K.; Center for Agricultural Research and Development; Chayanov, A.; Chenery, H.; Clark, J. B.; Cobweb theorem; Common property resources; Commons, R.; Computable general equilibrium models; Corn–hog cycle; Cournot, A. A.; Diminishing returns; Ely, R. T.; Environmental economics; European Association of Agricultural Economists; Ezekiel, M.; Factor mobility; Fisheries; General equilibrium; Georgescu-Roegen, N.; Griliches, Z.; Heady, E.;; Health; Hedonic prices; Hicks, J. R.; Horizontal integration; Human capital; Index numbers; Indicative planning; Industrial organization; Innovation; Input–output analysis; Johnson, D. Gale; Land; Land grant model; Laur School (Switzerland); Leontief, W.; Neoclassical economics; Neoclassical synthesis; Nourse, E.; Nutrition; Optimization theory; Physiocracy; Pollution; Poverty alleviation; Price stabilization; Production economics; Property rights; Quesnay, F.; Recursive models; Recursive programming; Rent; Ricardo, D.; Risk preferences; Samuelson, P. A.; Schultz, T. W.; Sering School (Germany); Serpieri School (Italy); Shackle, G. L. S.; Smith, A.; Spatial economics; Species loss; Stigler, G.; Stochastic programming methods; Subjective probability; Supply controls; Surplus production; Taylor, H. C.; technical change; Thünen, J. H. von; Treadmill hypothesis; Veblen, T.; Vertical integration; Warren, G.; Waugh, F.; Working, E.; Working, H.; Young, A

JEL Classifications

Q00

Agricultural economics arose in the late nineteenth century, combined the theory of the firm with marketing and organization theory, and developed throughout the twentieth century largely as an empirical branch of general economics. This emphasis was due to the historical importance of agriculture, and in the United States was made possible by the rich data compiled by the US Department of Agriculture beginning in the mid-nineteenth century. The discipline was closely linked to empirical applications of mathematical statistics and made early and significant contributions to econometric methods. From the 1960s on, as agricultural sectors in the OECD countries contracted, agricultural economists were drawn to the development problems of poor countries, to the trade and macroeconomic policy implications of agriculture in richer countries, and to a variety of issues in production, consumption, environmental and resource economics. This ramified the subject and enlarged its international focus, at the same time as its microeconomic, empirical and policy orientation distanced it from developments in general equilibrium theory, macroeconomic modelling, game theory and axiomatic social choice, which preoccupied many departments of economics throughout the late 20th century.

Retracing the evolution of agricultural economics, especially in the United States, requires an explanation of institutional innovation in 19th-century America (see Taylor and Taylor 1952). In the midst of the Civil War, President Lincoln created the Federal Department of Agriculture (later the US Department of Agriculture, USDA), empowered to collect a wide range of farm statistics. At the same time, legislation introduced by Vermont's Justin Morrill (previously blocked by the seceded South) was signed in 1862 by Lincoln. The Morrill Act established the Land Grant Colleges (financed through sales of government land) especially in the states of the Old Northwest Territory: Illinois, Indiana, Michigan, Ohio and Wisconsin. Their creation reflected both vast surpluses of land and the drive to improve plant and animal husbandry through applications of chemistry and biology. Eventually, the land grant model was replicated in every state as well as in some other countries. In 1887 the Hatch Act created the Agricultural Experiment Stations of USDA, which functioned together with the Land Grant Colleges to form a system of research, instruction and outreach to farmers (Cochrane 1993; Kerr 1987; Moore 1988). In 1914, extension education and outreach was formalized under the Smith–Lever Act. By the beginning of the 20th century, the application of scientific management to agricultural production created the foundations of the discipline.

Intellectual Origins

Agricultural economics in the United States derived from two intellectual streams. The first was neoclassical political economy and the theory of the firm applied to farm production. The second, borne of an economic crisis in American agriculture in the late 19th century, focused on strategies for organized marketing of agricultural commodities through collective bargaining and cooperatives. The first stream may be traced to the 18th-century Enlightenment and a preoccupation with land as a factor by the French Physiocrats. Francois Quesnay's *Tableau économique* (1758) organized a logical explanation of the conversion of land inputs to agricultural outputs and profit, anticipating modern production economics, input–output analysis and general equilibrium theory. His emphasis on surplus production was a touchstone of classical economics and exercised a direct influence over Adam Smith (Eltis 1975; Smith 1776, book II, chapter 9).

Like all 18th-century political economists, Smith could not ignore agricultural questions, even if he gave them less primacy than the Physiocrats. Together with Ricardo, Von Thünen and Malthus, he provided commentary on the difficulties of agricultural specialization, returns to land as a factor, issues of space and distance to market, and the long-run relation between arithmetic increases in food supply and geometric increases in demand due to population growth. Many pages of the *Wealth of Nations* dealt with agricultural questions, including the differential capacity for specialization and routinization of agriculture

versus industry and the arts of husbandry at the microeconomic level (1776, pp. 16, 143). Echoing the Physiocrats, Smith emphasized the central role of agriculture as a store of national wealth, and noted that compared with manufacturing, agriculture 'is much more durable, and cannot be destroyed by [the] violent convulsions' of war and political instability (1776, p. 427). In the same period, Arthur Young assembled comprehensive data on production, rents and land tenure in Great Britain. Serving as editor of the *Annals of Agriculture* from 1768 to 1770, he collected his data and observations into nine volumes of 4,500 pages, which have proved to be of continuing value especially to economic historians (for example, Allen 1992). Ricardo (1821, p. 44) was famously concerned with returns to land as a fixed factor 'for the use of the original and indestructible powers of soil'. He also distinguished between productivity enhancements due to augmentation of the soil and improvements in machinery and the capitalization of various investments or policies (such as taxes) into the value of land (1821, pp. 57–61; 246). Von Thünen's (1828) analysis of the extensive margin and the relationship between distance to market and rent made him, in Marshall's view, the first agricultural economist among economists, who with Cournot provided the inspiration for marginalist economics (Day and Sparling 1977, p. 93).

It was the neoclassical developments of the late 19th century, however, that provided the main foundations for agricultural economics. Marshall's *Principles* (1890) first clearly established the link from diminishing marginal utility in exchange to decreasing marginal productivity on the supply side. Veblen (1900) dubbed Marshall's work 'neoclassical' to distinguish it from classical labour theories of value. The elaboration of Marshall's theory of the firm, and attempts to measure and statistically validate the relationship between input costs, output prices, and farm profits distinguished agricultural economics well into the 20th century, and linked it firmly to the neoclassical syntheses of Hicks (1939) and Samuelson (1947).

To this was added a second stream of marketing and organizational issues growing out of the extended farm depression from the 1870s to the 1890s. Joined with labour interests, farmers sought marketing outlets and modes of organization that would give them greater bargaining power, notably cooperatives popular in northern Europe and Scandinavia, where many recently arrived American farmers originated (Jesness 1923). Even after the business cycle turned upward after 1897, the Land Grant colleges emphasized farm management. The result was the organization in 1910 of the American Farm Management Association. Farm managers were focused on the physical, technical and scientific aspects of production, especially the new field of agronomy.

Many early agricultural economists regarded farm management as a sub-field, and agricultural economics as an applied version of general economics. Beginning in 1907, at the tenth American Economic Association (AEA) meetings, a session was devoted to 'What is agricultural economics?' Thereafter, the AEA regularly included sessions on the economics of agriculture. In 1915 the National Association of Agricultural Economists was formed. In 1917 the AEA meeting was held jointly with the National Agricultural Economics Association and the American Farm Management Association, and talks began on a merger of the latter two. This was realized in 1919 in the form of the American Farm Economics Association, with Henry C. Taylor of the University of Wisconsin as President (Taylor 1922; Cochrane 1983). It retained this title until 1968, when it became the American Agricultural Economics Association (AAEA).

The Discipline Expands

As Cochrane (1983, p. 66) observed, 'the first flowering of agricultural economics as an applied field of economics occurred at the University of Wisconsin in the period of 1900–1920. The second flowering occurred at the University of Minnesota in the period of 1918–1928.' A department of agricultural economics was established at Wisconsin in 1909 by Henry C. Taylor and colleagues such as Benjamin Hibbard. Taylor's text,

An Introduction to the Study of Agricultural Economics (1905), applied Marshallian principles to farm production, and developed production functions showing increasing, steady and diminishing returns. Among the most influential leaders in the young subject was Taylor's student at Wisconsin, John D. Black, who also studied under John R. Commons and Richard T. Ely (who himself authored an influential, though unpublished, 1904 study on the economics and property rights of irrigation). Their emphasis on land and institutions permeated the discipline and was reflected in the journal *Land Economics*, which began publication at Madison in 1925.

Black, a follower of Marshall and John Bates Clark, received his Ph.D. in 1918 and moved to the University of Minnesota, where he remained a dominant force until hired by Harvard in 1927. By the mid-1920s Black's leadership had marked him, together with George F. Warren of Cornell and Edwin G. Nourse of Iowa State, as 'the most influential economist in the United States dealing with the problems of agriculture' (Galbraith 1959, p. 10). Together with a cadre of other young economists working with the Bureau of Agricultural Economics (BAE), created in USDA in 1921, Black set the tone for research in the field from the 1920s until the advent of the Second World War.

Black's text, *Introduction to Production Economics* (1926), became the standard. His emphasis on the theory of the firm was complemented by his colleague Holbrook Working's econometric explorations. Working's 1922 bulletin, 'Factors Determining the Price of Potatoes in St. Paul and Minneapolis', was among the first to derive an empirical demand curve (H. Working 1922, 1925). It was followed by his brother E. J. Working's widely cited 1927 article, 'What Do Statistical "Demand Curves" Show?' The Workings and colleague Warren Waite continued to expand research into price analysis in the interwar years. Minnesota's Frederick V. Waugh contributed the first quantitative study of quality characteristics as determinants of prices, recognized as a forerunner of hedonic price analysis. Appearing as 'Quality Factors Influencing Vegetable Prices' (1928), it noted that if 'a premium for certain qualities and types of products is more than large enough to pay the increased cost of growing a superior product, the individual can and will adapt his production and marketing policies to market demand' (quoted in Berndt 1991, p. 106).

Taylor, Black, Warren and Nourse were followed by a group of young empiricists and econometricians who continued to develop the USDA Bureau of Agricultural Economics (BAE). Tolley et al. (1924) showed how production surfaces in three dimensions could express diminishing returns to inputs, a concept readily grasped by agricultural field scientists. They then derived cost surfaces showing the relationship between costs, relative prices, and profit maximization. Ezekiel followed this empirical work with his 1930 volume *Methods of Correlation Analysis*, which became a standard text on regression analysis, and in 1938 with a state-of-the-art description of cobweb and recursive models illustrated by the corn–hog cycle. Leontief (1971, p. 5) would call this and other early agricultural economists' work 'An exceptional example of a healthy balance between theoretical and empirical analysis .. .' and 'the first among economists to make use of the advanced methods of mathematical statistics'.

By the 1930s departments of agricultural economics had been established in many US universities, where technical and institutional issues affecting agricultural production formed the core subjects. In addition to the leading roles played by Cornell, Illinois, Iowa State, Minnesota, Purdue and Wisconsin, a major research programme was established at the University of California-Berkeley (and a later campus at Davis) with the endowment of the Giannini Foundation. At Iowa State, future Nobel Laureate T.W. Schultz arrived in 1930 with a Ph.D. from Wisconsin, and then served as department head from 1934 to 1943 until leaving for Chicago. Schultz attracted numerous talents including Kenneth Boulding, George Stigler, D. Gale Johnson and Earl O. Heady, several of whom would also leave for Chicago following controversy surrounding oleomargarine and the Iowa butter industry (Beneke 1998). The butter–margarine dispute was typical of agricultural economists' conflicts with interest

groups in a profession seldom sheltered from political winds, especially at state universities. Partly for this reason, several private universities also made substantial contributions to agricultural economics research. In addition to Black (and later Galbraith) at Harvard, the University of Chicago remained a center of research excellence. At Vanderbilt, Nicholas Georgescu-Roegen, a demand theorist and econometrician, expressed path-breaking insights into the physical process underlying economic activity, and contributed a deep critique of agrarianism and Marxian misunderstandings of agricultural production (Georgescu-Roegen 1960).

Earl O. Heady remained at Iowa State, creating a post-war engine of applied research, the Center for Agricultural Research and Development (CARD), in 1957. He pioneered the application of programming methods first developed for war planning, analysing how inputs could most efficiently be employed in producing agricultural outputs. This made the discipline a centre for research in applications of optimization theory. Heady authored or oversaw hundreds of mainly empirical production studies, exemplified by Heady and Dillon (1961) and Heady and Candler (1958). He also pioneered the application of computing power to problem-solving in applied economics. This included work on human and animal diet rations and consumption (for example, Waugh 1951; Heady 1951). Farm management also saw optimization applications in work by Hildreth (1957a) among others. By the late 1950s Bellman's dynamic programming principle was applied to optimal wheat rotations by Burt and Allison (1963). Agricultural economics also began to grapple empirically with uncertainty through stochastic programming methods, including Hildreth's (1957b) work and Hazell's applications (Hazell 1971). French economists Boussard and Petit applied Shackle's 'focus loss' concept of uncertainty to agriculture (1967). The application of subjective probability concepts to agriculture was surveyed by Dillon (1971) and Anderson et al. (1977).

Yet another outgrowth of optimization theory was analysis of the growth and decline of farms in modern economies, including contributions by German agricultural economists Heidhues (1966) and De Haen (De Haen and Heidhues 1973). Behavioural adjustment ('supply response') in agriculture was studied using recursive programming models (Henderson 1959), and generalized by Day (1963), following the path set by Nerlove (1958). Optimal storage rules were analysed by Gustafson (1958). Spatial issues in agriculture analysed best-location decisions (Egbert and Heady 1961), and interregional supply–demand equilibrium issues (for example, Fox 1953). An extensive bibliography of spatial and temporal equilibrium models was published by Judge and Takayama (1973).

New Frontiers

Two additional applications of optimization theory pushed agricultural economics in the 1960s and 1970s toward new frontiers: natural resources and agricultural development in developing countries. These helped attract a new generation of economists concerned less with domestic farm production than with environmental issues and poverty alleviation in the Third World. Natural resources were analysed as problems of materials shortages and treated as a form of capital, following the early analytical leads of Hotelling (1931) and Ciriacy-Wantrup (1952). Especially after the Paley Commission Report of 1952 led to the creation of Resources for the Future in Washington, DC, a new group of economists applied themselves to these issues. Fisheries were studied by Scott (1955) and Crutchfield and Zellner (1962); groundwater allocation over time was considered as a dynamic programme with stochastic state variables in a series of articles by Burt (for example, Burt 1966; Burt and Cummings 1970). These dynamic models were extended to interregional investments in water in studies such as Cummings and Winkelmann (1970). By the 1970s, environmental pollution became a major subject of applied economics, pulling many in the profession away from a restricted view of agricultural issues as matters of yields and production in acknowledgement of the sector's negative external effects and market failures.

Agricultural development in developing countries, meanwhile, was an important area of applied economics in project evaluation, supported by multilateral and bilateral aid agencies such as the World Bank, the Food and Agriculture Organization of the UN (FAO) and US Agency for International Development. At Stanford, the Food Research Institute (1921–1995) established an internationally focused research programme. The development problem in the Third World was seen largely as an imbalance between agricultural and manufacturing sectors, with a need to right this balance by drawing low-productivity resources out of agriculture (Lewis 1954; Mellor 1966; Timmer 2002). Hollis Chenery at the World Bank exemplified the analysis of agriculture's sectoral role (Chenery and Syrquin 1975). However, unlike the United States and some other OECD countries, data limitations in poor countries restricted the early application of optimization models at the microeconomic level. Indeed, T.W. Schultz's famous *Transforming Traditional Agriculture* (1964) relied mainly on stylized representations of 'rational but poor' farmers and descriptive analysis from anthropologists.

Throughout the 1950s and 1960s the agricultural sector continued to contract in the OECD countries, setting the tone for policy debates. Many agricultural economists saw the 'farm problem' as one of surplus labour supplying farm commodities in excess of domestic demand. Analysing low agricultural prices as a matter of chronic oversupply, aggravated by rapid technological improvements and productivity gains in the face of inelastic demand, Cochrane (1958) proposed his treadmill hypothesis: rapid and early adopters of productivity-improving technology will reap the lion's share of rents to innovation, as laggards are forced off the farm, while Brewster (1959) considered the social and policy implications of these trends. In the early 1960s, serving as presidential adviser, Cochrane advocated a solution to excess production in the form of federally mandated supply control. When it became clear that the major commodity groups would vote down the enabling referenda, and that its success would raise prices to consumers, President Kennedy abandoned the scheme. Thereafter, although mandated supply control retained adherents (not including Cochrane), US agricultural policy shifted towards exports as a vent-for-surplus.

This opened the way to consideration of agriculture in an open economy, and a new policy emphasis on the macroeconomics of the food sector (Schuh 1974, 1976; Cochrane and Runge 1992; Ardeni and Freebairn 2002; Abbott and McCalla 2002). In the 1980s, this open economy analysis was supported by the development of large-scale computable general equilibrium models linking agriculture to trade (for example, Hertel 1997) as well as more traditional macroeconomic sectoral forecasting models (for example, Myers et al. 1987). Together, the large-scale models allowed alternative trade and agricultural policy approaches to be simulated and compared to the *status quo* (for example, Cochrane and Runge 1992).

International Reach

The intellectual antecedents of agricultural economics make clear that the field has never been restricted to the United States. In 1905, the International Agricultural Institute was founded in Rome as the forerunner of the FAO. In Great Britain, an Agricultural Economics Research Institute was established at Oxford in 1913, and in 1945 became part of the School of Rural Economy, merging with Queen Elizabeth House and the Institute for Commonwealth Studies in 1986. Oxford led the creation of the International Association of Agricultural Economists and helped coordinate its first conference in 1929 at Dartington Hall, Devon and a second in 1930 at Cornell. These were largely Anglo-American meetings, although by the third meeting in Germany in 1934, 19 different countries were represented. At Cambridge, a Department of Estate Management was transformed into a Department of Land Economy in the 1960s. At Wye, an agricultural college was founded in 1894. The college was awarded a royal charter in 1948 and in 2000 its agricultural economics department became part of Imperial College London.

On the Continent, followers of Von Thünen had developed marginalist principles and farm accounting methods in the late 19th and early 20th century represented by the Laur School in Switzerland and the Sering and Serpieri Schools in Germany and Italy. However, their capacity was limited by poor data, few marketing studies, and a weak connection to production economics (Nou 1967; Raeburn and Jones 1990, p. 13). In 1948 a French professional association began, and a Department of Agricultural Economics was created at the Institut National de la Recherche Agronomique (INRA) in 1955 (Petit 1982). A European Association of Agricultural Economists was founded in 1975 in Uppsala, Sweden. By the late 1980s, it was estimated that 3,000–5,000 European professionals were engaged in full-time agricultural economics research dispersed in hundreds of research institutes, universities and government offices (Hanf 1988). Among the leaders were the French government's INRA, the Universities of Goettingen and Kiel in Germany, the University of Padova in Italy, Wageningen University in the Netherlands, and the aforementioned activities in Great Britain.

In Canada, agricultural economics began at the Ontario Agricultural College (now the University of Guelph) in 1907. Noteworthy research departments of agricultural economics were established at the University of Guelph, Ontario, McGill University in Montreal, Laval University in Quebec, and the Universities of Manitoba, Alberta, Saskatchewan and British Columbia.

The Australian Agricultural Economics Society was founded in Sydney in 1957, following the models of the US, British and Canadian associations. In 1975, a New Zealand branch of the association was established at a meeting in Christchurch. The leading Australian institution in creating a separate department was the University of New England at Armidale, which in 1958 began a 4-year course. Supported by grants from the Commonwealth Bank, a chair of agricultural economics was appointed at the University of Sydney in 1951 (Campbell 1985). While maintaining the specialty within economics rather than a separate department, major research was also undertaken beginning in the 1950s and 1960s at the University of Adelaide and at the University of Melbourne, and later at the Australian National University in Canberra and the University of Western Australia in Perth. All of these universities were closely linked to the national Bureau of Agricultural Economics (BAE), which became the Australian Bureau of Agriculture and Resource Economics (ABARE) in 1987 (Miller 1985).

In Russia, interest in agricultural economics may be traced to the establishment in 1865 of the Moscow Agricultural Academy. In 1929 Lenin created the Russian Academy of Agricultural Sciences, following conflicts between Chayanov and Marxist agriculturalists. After Stalin's rise to power in 1930, agricultural research was fully politicized with well-known results, including the purge of many academic researchers (Nazarenko 2004). In the 1950s, concepts such as profit and cost were revived, and central planners embraced modelling and forecasting. Since the 1990s, agricultural reforms have led to dissension in the Russian discipline (Klyukach 2004).

In Brazil, the Rockefeller and Ford Foundations and the US Agency for International Development provided core support for agricultural economics research, beginning in the late 1950s. Four US universities were directly involved: Purdue, Wisconsin, Ohio State and Arizona.

In India, a Society of Agricultural Economics was established in 1939. The advent of indicative economic planning in the 1950s stimulated analytical studies to assist in the Plan. Due to the overwhelming importance of agriculture as a supplier of wage goods, the sector attracted considerable analysis, in which Indian agricultural universities, established on the land-grant model, consciously borrowed methods from their US counterparts, notably Earl O. Heady and the CARD group at Iowa State (Bhide 1994, p. 119).

In China, missionary efforts to promote agricultural research and development by the Presbyterian Church of New York during the first quarter of the 20th century resulted in a Cornell University–University of Nanking collaboration, led beginning in 1914 by John Lossing Buck (1973). J. L. Buck's contributions included early agricultural surveys and analysis of Communist production into the 1960s (Buck 1943; Buck et al. 1966).

Late 20th Century

Since the 1970s, seven broad subjects have defined the most distinctive contributions of agricultural economics: technical change and the returns to human capital investments; environmental and resource issues; trade and economic development; agricultural risk and uncertainty; price determination and income stabilization; market structure and the organization of agricultural businesses; and consumption and food supply chains.

The study of technical change, innovation and returns to investments in human capital in agriculture attracted some of the most talented economists of the post-war generation, such as Zvi Griliches (1957, 1958, 1963, 1964). Anticipating debates among economic growth theorists over 'embodied' technical change due to improvement in the quality of capital inputs (versus 'disembodied' changes without new net capital investments), Cochrane (1953) criticized Schultz (1953) for failing to account for capital requirements in agriculture and a resulting overemphasis on weather variations in describing growth in yields. Focusing on the direction of agricultural innovation, Ruttan (1956) and Hayami and Ruttan (1971) emphasized the Hicks-non-neutrality of technical change in both labour-saving US and land- saving Japanese agriculture. This approach was extended in a formal framework by Binswanger (1974). Based on Hicks's (1932) analysis of relative factor prices as the inducement to alternative paths of innovation, the induced innovation argument was extended into an explanation of priority setting by public sector agencies, leading research towards abundant factor use that lowered social costs of production (Peterson and Hayami 1977, p. 504). How to measure productivity and technical change in agriculture using alternative index numbers attracted both theorists and applied econometricians (for example, Jorgenson and Griliches 1967; Lau and Yotopoulos 1971). Finally, analysts considered the welfare gains and losses resulting from farm mechanization (Schmitz and Seckler 1970).

Agricultural economists also delved into the role of productivity embodied in labour as 'human capital', a natural reflection of the huge public investments in research and education by the US land grant system. Surveyed by T. W. Schultz (1971), this line of research attracted work by Peterson (1969), Huffman (1974) and general economists such as Nelson and Phelps (1966), and led to widening emphasis on private and social returns to research including Peterson (1967), Evenson (1967), Evenson and Kislev (1976) and Alston et al. (2000). It also led to analysis of how research ought to be organized in order to maximize its aggregate benefits. Alston et al. (1998) developed a comprehensive summary of this priority- setting problem (see Huffman 2002; Sunding and Zilberman 2002).

Environmental and resource issues, as noted, became a significant focus of the profession in the 1970s and beyond, partly in recognition of the pollution and species losses resulting from modern agricultural systems. Surveyed by Lichtenberg (2002), the economics of agriculture and the environment analysed the perverse incentives created by agricultural subsidies and the agency problems of monitoring agricultural practices (for example, Chambers and Quiggin 1996; Just and Antle 1990; Segerson 1988). Induced innovation theory was broadened to explain how technical innovations such as irrigation might give rise to new water quality issues and thus new institutional responses (for example, Runge 1987; Caswell et al. 1990). Apart from specific agriculture–environment interactions, resource economists emphasized the critical role of property rights in the use and management of resources, especially those held publicly or in common, notably in developing countries (Runge 1981; Bromley 1991; Walker et al. 2000).

Trade and development also dominated agricultural economics research, especially after the mid-1980s, as global trade negotiations increasingly hinged on struggles between heavily subsidized farm sectors in OECD countries and the highly taxed sectors of the developing world (Anderson and Hayami 1986; Kreuger et al. 1991–1992; Sumner and Tangermann 2002). An overview of post-war agricultural trade policy was given by D. G. Johnson (1977); a synthetic treatment of agriculture–trade interactions was provided by Karp and Perloff (2002). Meanwhile, a

major share of agricultural economics literature was devoted to microeconomic studies of agricultural change and food insecurity in developing countries, and to macroeconomic linkages with other sectors and global trade (for example, Barrett 2002; Runge et al. 2003).

Risk and uncertainty are inherent in agriculture and their relevance has drawn interest from many agricultural economists, especially in developing-country decision environments (see Moschini and Hennessey 2002). Roumasset (1976) conducted an early assessment of risk aversion and the adoption of hybrid rice in the Philippines. Dillon and Scandizzo (1978) analysed risk preferences among small farmers in Brazil, while Moscardi and de Janvry (1977) analysed Mexican maize production and the response to risk. Antle (1987) and Myers (1989) provided econometric tests for risk aversion by farmers while Goodwin and Smith (1995) and Miranda and Glauber (1997) considered why crop insurance contracts fail effectively to pool risk without reinsurance.

Price determination and stabilization of agricultural prices as a focus of research arose as a direct consequence of widespread instability in agricultural commodities markets. Tomek and Robinson (1977) surveyed the post-war literature through the 1970s, including the analysis of Cochrane (1958) and Gray and Rutledge (1971). In response to widespread calls for buffer stocks and other mechanisms to affect prices countercyclically, Newbury and Stiglitz (1981) offered a comprehensive (and sceptical) assessment of the advantages of stabilization policy. A more recent survey was developed by Wright (2002).

The organizational structure of farms and the role of economies of scale, scope, technological change, capital and labour mobility were reviewed by Chavas (2002). Farm size was analysed as a function of the opportunity cost of labour and the price of machinery (Kislev and Peterson 1982). Farm structure and the economics of contracting was also an additional area of risk and agency studies (Allen and Lueck 1998; Hueth and Ligon 2001; Knoeber and Thurman 1995). Despite their declining importance in many rural markets, cooperatives continued to attract analysis (for example, Sexton 1990).

A final area of broad interest was food consumption and supply chains in the food industry. Taking an industrial organization approach, Sexton and Lavoie (2002) provided an overview, emphasizing vertical and horizontal integration and imperfect competition as forces driving the sector, with implications for consumer choice, nutrition and health.

In the 21st century, the profession has continued to reach beyond the agricultural sector, expanding its scope through numerous applications of relevant economic theory. Meanwhile, the high level of abstraction in economics characteristic of the last half of the 20th century appears to have given way to new interest in empirical and experimental studies, suggesting that the distance between agricultural economics and its mother discipline may narrow in the years ahead.

See Also

- ▶ Agriculture and Economic Development
- ▶ Econometrics

Bibliography

Note: This article can only gesture to the agricultural economics literature. The reader is referred to L. R. Martin (general editor) and the four-volume *Survey of Agricultural Economics Literature*, Minneapolis, University of Minnesota Press, 1977–1987, and Wallace C. Olsen et al., *Agricultural Economics and Rural Sociology: The Contemporary Core Literature*, Ithaca: Cornell University Press, 1991. A compilation of analytical and interpretive essays is B. L. Gardner and G. C. Rausser (eds.), *Handbook of Agricultural Economics*, vols. 1A, 1B, 2A and 2B, Amsterdam: North-Holland, 2002. An internet-based open source of information is AgEcon Search, maintained at the University of Minnesota: http://www.apec.umn.edu/AgEcon.html (accessed 16 August 2006).

Abbott, P., and A. McCalla. 2002. Agriculture in the macroeconomy. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, vol. 2A. Amsterdam: North- Holland.

Allen, R. 1992. *Enclosure and the yeoman*. Oxford: Clarendon Press.

Allen, D., and D. Lueck. 1998. The nature of the farm. *Journal of Law and Economics* 41: 343–386.

Alston, J., G. Norton, and P. Pardey. 1998. *Science under scarcity: Principles and practices for agricultural*

evaluation and priority setting. New York: CAB International.

Alston, J., C. Chan-Kang, M. Marra, P. Pardey, and T. Wyatt. 2000. *A meta- analysis of rates of return to agricultural R&D* . Washington, DC: International Food Policy Research Institute.Research Report No. 113

Anderson, K., and Y. Hayami. 1986. *The political economy of agricultural protection: East Asia in international perspective*. London: Allen and Unwin.

Anderson, J., J. Dillon, and B. Hardaker. 1977. *Agricultural decision analysis*. Ames: Iowa University Press.

Antle, J. 1987. Econometric estimation of risk producers' risk attitudes. *American Journal of Agricultural Economics* 69: 509–522.

Ardeni, P., and J. Freebairn. 2002. The macroeconomics of agriculture. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, vol. 2A. Amsterdam: North-Holland.

Barrett, C.B. 2002. Food security and food assistance. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, vol. 2B. Amsterdam: North-Holland.

Beneke, R. 1998. T.W. Schultz and pamphlet No. 5: The oleo margarine war and academic freedom. *Choices* 2nd Quarter, 4–8.

Berndt, E. 1991. *The practice of econometrics: Classical and contemporary.* Reading: Addison-Wesley.

Bhide, S. 1994. The science of agricultural economics in India: Implications of Heady's contributions. In *Earl O. Heady: His impact on agricultural economics*, ed. J. Langley, G. Vocke, and L. Whiting. Ames: Iowa State University Press.

Binswanger, H. 1974. A microeconomic approach to induced innovation. *Economic Journal* 84: 940–958.

Black, J. 1926. *Introduction to production economics*. New York: Henry Holt and Co..

Boussard, J., and M. Petit. 1967. Representation of farmers' behavior under uncertainty with a focus loss constraint. *Journal of Farm Economics* 49: 869–880.

Brewster, J. 1959. The impact of technical advance and migration on agricultural society and policy. *Journal of Farm Economics* 41: 1169–1184.

Bromley, D. 1991. *Environment and economy: Property rights and public policy.* Oxford: Blackwell.

Buck, J. 1943. *An agricultural survey of Szechwan province*. Chungking: Farmers Bank of China.

Buck, J. 1973. *Development of agricultural economics at the University of Nanking, Nanking, China 1920–1946*, Agricultural Development Bulletin No. 25. Ithaca: Cornell University.

Buck, J., O. Dawson, and Y. Wu. 1966. *Food and agriculture in communist China*. New York: Praeger.

Burt, O. 1966. Economic control of groundwater reserves. *Journal of Farm Economics* 48: 632–647.

Burt, O., and J. Allison. 1963. Farm management decisions with dynamic programming. *Journal of Farm Economics* 45: 121–136.

Burt, O., and R. Cummings. 1970. Production and investment in natural resource industries. *American Economic Review* 60: 576–590.

Campbell, K. 1985. Some reflections on the development of agricultural economics in Australia. In *Agricultural economics in Australia*, ed. S. Bearman. Armidale: Department of Agricultural Business and Management, University of New England.

Caswell, M., E. Lichtenberg, and D. Zilberman. 1990. The effects of pricing policies on water conservation and drainage. *American Journal of Agricultural Economics* 72: 883–890.

Chambers, R., and J. Quiggin. 1996. Non-point-source Pollution Regulation as a Multi-task principal–agent problem. *Journal of Public Economics* 59: 95–116.

Chavas, J. 2002. Structural change in agricultural production, economics, technology and policy. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, vol. 1A. Amsterdam: North-Holland.

Chenery, H., and M. Syrquin. 1975. *Development, 1950–1970*. Oxford: Oxford University Press.

Ciriacy-Wantrup, S. 1952. *Resource conservation – economics and policies*. Berkeley: University of California Press.

Cochrane, W. 1953. Professor Schultz discovers the weather. *Journal of Farm Economics* 35: 280–283.

Cochrane, W. 1958. *Farm prices: Myth and reality.* Minneapolis: University of Minnesota Press.

Cochrane, W. 1983. *Agricultural economics at the University of Minnesota 1886–1979,* Miscellaneous publication no. 21. Institute of Agriculture, Forestry and Home Economics, University of Minnesota.

Cochrane, W. 1993. *The development of American agriculture: A historical analysis*. Minneapolis: University of Minnesota Press.

Cochrane, W., and C. Runge. 1992. *Reforming farm policy: Toward a national agenda*. Ames: Iowa State University Press.

Crutchfield, J., and A. Zellner. 1962. *Economic aspects of the Pacific halibut industry*, Fishery Industrial Research No. 1. Washington, DC: United States Department of the Interior.

Cummings, R., and D. Winkelmann. 1970. Water resource management in arid environs. *Water Resources Research* 6: 1559–1568.

Day, R. 1963. *Recursive programming and production response*. Amsterdam: North-Holland.

Day, R., and E. Sparling. 1977. Optimization models in agricultural and resource economics. In *A survey of agricultural economics literature*, ed. L. Martin, vol. 2. Minneapolis: University of Minnesota Press.

De Haen, H., and T. Heidhues. 1973. *Recursive programming models to simulate agricultural development – applications in West Germany*, Working Paper No. 18. Göttingen: Institute for Agricultural Economics.

Dillon, J. 1971. An expository review of Bernoullian decision theory in agriculture: Is utility futility? *Review of Marketing and Agricultural Economics* 39: 3–80.

Dillon, J., and P. Scandizzo. 1978. Risk attitudes of subsistence farmers in Northeast Brazil: A sampling approach. *American Journal of Agricultural Economics* 60: 425–435.

Egbert, A., and E. Heady. 1961. *Regional adjustments in grain production: A linear programming analysis*, Technical Bulletin No. 1241. Washington, DC: USDA.

Eltis, W. 1975. François Quesnay: A reinterpretation. 1. The Tableau Economique. *Oxford Economic Papers* 27: 167–200.

Evenson, R. 1967. The contribution of agricultural research to production. *Journal of Farm Economics* 49: 1415–1425.

Evenson, R., and Y. Kislev. 1976. A stochastic model of applied research. *Journal of Political Economy* 84: 265–281.

Ezekiel, M. 1930. *Methods of correlation analysis*. New York: Wiley.

Ezekiel, M. 1938. The cobweb theorem. *Quarterly Journal of Economics* 52: 255–280.

Fox, K. 1953. A spatial equilibrium model of livestock-feed economy in the United States. *Econometrica* 21: 547–566.

Galbraith, J. 1959. John D. Black: A portrait. In *Economics for agriculture*, ed. J. Cavin. Cambridge, MA: Harvard University Press.

Georgescu-Roegen, N. 1960. Economic theory and agrarian economics. *Oxford Economic Papers* 12: 1–40.

Giles, A. 2000. The AES, 1926–2001: A view from the archives. *Journal of Agricultural Economics* 52: 1–19.

Goodwin, B., and V. Smith. 1995. *The economics of crop insurance and disaster aid*. Washington, DC: AEI Press.

Gray, R., and D. Rutledge. 1971. The economics of commodity futures markets: a survey. *Review of Marketing and Agricultural Economics* 39: 57–108.

Griliches, Z. 1957. Hybrid corn: An exploration in the economics of technological change. *Econometrica* 25: 501–522.

Griliches, Z. 1958. Research costs and social returns: Hybrid corn and related innovations. *Journal of Political Economy* 66: 419–431.

Griliches, Z. 1963. The sources of measured productivity growth: United States agriculture, 1940–1960. *Journal of Political Economy* 71: 331–346.

Griliches, Z. 1964. Research expenditures, education, and the aggregate agricultural production function. *American Economic Review* 54: 961–974.

Gustafson, R. 1958. Implications of recent research on optimal storage rules. *Journal of Farm Economics* 40: 290–300.

Hanf, C. 1988. Adjustment necessities in European agricultural economics research. *European Review of Agricultural Economics* 15: ix–xvi.

Hayami, Y., and V. Ruttan. 1971. *Agricultural development: An international perspective*. Baltimore: Johns Hopkins University Press.

Hazell, P. 1971. A linear alternative to quadratic and semivariance programming for farm planning under uncertainty. *American Journal of Agricultural Economics* 53: 53–62.

Heady, E. 1951. A production function and marginal rates of substitution in the utilization of feed resources by dairy cows. *Journal of Farm Economics* 33: 485–498.

Heady, E., and W. Candler. 1958. *Linear programming methods*. Ames: Iowa State University Press.

Heady, E., and J. Dillon. 1961. *Agricultural production functions*. Ames: Iowa State University Press.

Heidhues, T. 1966. A recursive programming model of farm growth in northern Germany. *Journal of Farm Economics* 48: 668–684.

Henderson, J. 1959. The utilization of agricultural land: a theoretical and empirical inquiry. *The Review of Economics and Statistics* 41: 242–259.

Hertel, T., ed. 1997. *Global trade analysis modeling and applications*. New York: Cambridge University Press.

Hicks, J. 1932. *The theory of wages*. London: Macmillan.

Hicks, J. 1939. *Value and capital: An inquiry into some fundamental principles of economic theory.* Oxford: Clarendon Press.

Hildreth, C. 1957a. Problems of uncertainty in farm planning. *Journal of Farm Economics* 39: 1430–1441.

Hildreth, C. 1957b. Some problems and possibilities of farm programming. In *Fertilizer innovations and resource use*, ed. E. Baum et al. Ames: Iowa State College Press.

Hotelling, H. 1931. The economics of exhaustible resources. *Journal of Political Economy* 39: 137–175.

Hueth, B., and E. Ligon. 2001. Agricultural markets as relative performance evaluation. *American Journal of Agricultural Economics* 83: 318–328.

Huffman, W. 1974. Decision making: the role of education. *American Journal of Agricultural Economics* 56: 85–97.

Huffman, W. 2002. Human capital, education and agriculture. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, vol. 1A. Amsterdam: North- Holland.

Jesness, O. 1923. *The cooperative marketing of farm products*. Philadelphia: Lippencott.

Johnson, D. 1977. Postwar policies relating to trade in agricultural products. In *A survey of agricultural economics literature*, ed. L. Martin. Minneapolis: University of Minnesota Press.

Jorgenson, D., and Z. Griliches. 1967. The explanation of productivity change. *Review of Economic Studies* 34: 249–283.

Judge, G., and T. Takayama. 1973. *Studies in economic planning over space and time*. Amsterdam: North–Holland.

Just, R., and J. Antle. 1990. Interactions between agricultural and environmental Policies: A conceptual framework. *American Economic Review* 80: 197–202.

Karp, L., and J. Perloff. 2002. A synthesis of agricultural trade economics. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, vol. 2B. Amsterdam: North-Holland.

Kerr, N. 1987. *The legacy: A centennial history of the state agricultural experiment stations 1887–1987*. Columbia: Missouri Agricultural Experiment Station.

Kislev, Y., and W. Peterson. 1982. Prices, technology and farm size. *Journal of Political Economy* 90: 578–595.

Klyukach, V. 2004. Stages in the formation and modern development of Russian agricultural economics.

Ekonomika Sel'skokhozyaistvennykh I Pererabatyvayushchikh Predpriyatii 5: 10–13.

Knoeber, C., and W. Thurman. 1995. 'Don't Count Your Chickens... ': Risk and risk shifting in the broiler industry. *American Journal of Agricultural Economics* 77: 486–496.

Kreuger, A., Schiff, M. and Valdes, A. 1991–94. *The political economy of agricultural pricing policy*, *4 vols*. Baltimore: Johns Hopkins University Press.

Lau, L., and P. Yotopoulos. 1971. A test for relative efficiency and application to Indian agriculture. *American Economic Review* 61: 94–109.

Leontief, W. 1971. Theoretical assumptions and nonobserved facts. *American Economic Review* 61: 1–7.

Lewis, W. 1954. Economic development with unlimited supplies of labour. *The Manchester School* 22: 3–42.

Lichtenberg, E. 2002. Agriculture and the environment. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, vol. 2A. Amsterdam: North-Holland.

Marshall, A. 1890. *Principles of economics*. 1st ed. London: Macmillan and Co..

Mellor, J. 1966. *The economics of agricultural production*. Ithaca: Cornell University Press.

Miller, G. 1985. Agricultural economics in Canberra. In *Agricultural economics in Australia*, ed. S. Bearman. Armidale: Department of Agricultural Business and Management, University of New England.

Miranda, M., and J. Glauber. 1997. Systematic risk, reinsurance, and the failure of crop insurance markets. *American Journal of Agricultural Economics* 79: 206–215.

Moore, G. 1988. The involvement of experiment stations in secondary agricultural education, 1887–1917. *Agricultural History* 62: 164–176.

Moscardi, E., and A. de Janvry. 1977. Attitudes toward risk among peasants: An econometric approach. *American Journal of Agricultural Economics* 59: 710–716.

Moschini, G., and D. Hennessey. 2002. Uncertainty, risk aversion and risk management for agricultural producers. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, vol. 2B. Amsterdam: North-Holland.

Myers, R. 1989. Econometric testing for risk averse behavior in agriculture. *Applied Economics* 21: 542–552.

Myers, W.H., A. Womack, S. Johnson, J. Brandt, and R. Young. 1987. Impacts of alternative programs indicated by FAPRI analysis. *American Journal of Agricultural Economics* 69: 972–979.

Nazarenko, V. 2004. Conflict in Soviet agricultural economics in the 1920s and 1930s. *Ekonomika Sel'skokhozyaistvennykh I Pererabatyvayushchikh Predpriyatii* 6: 8–11.

Nelson, R., and E. Phelps. 1966. Investment in humans, technological diffusion, and economic growth. *American Economic Review* 56 (2): 69–75.

Nerlove, M. 1958. *The dynamics of supply, estimation of farmers response to price*. Baltimore: Johns Hopkins University Press.

Newbury, D., and J. Stiglitz. 1981. *The theory of commodity price stabilization – a study in the economics of risk*. Oxford: Clarendon.

Nou, J. 1967. *Studies in the development of agricultural economics in Europe*. Uppsala: Almquist and Wiksells.

Olsen, W., et al. 1991. *Agricultural economics and rural sociology: The contemporary core literature*. Ithaca: Cornell University Press.

Peterson, W. 1967. Return to poultry research in the United States. *Journal of Farm Economics* 49: 656–669.

Peterson, W. 1969. The allocation of research, teaching and extension personnel in U.S. colleges of agriculture. *American Journal of Agricultural Economics* 51: 41–56.

Peterson, W., and Y. Hayami. 1977. Technical change in agriculture. In *A survey of agricultural economics literature*, ed. L. Martin, vol. 1. Minneapolis: University of Minnesota Press.

Petit, M. 1982. Is there a french school of agricultural e? *Journal of Agricultural Economics* 23: 325–337.

Quesnay, F. 1758. *The economical table*. Milton Keynes: Lightning Source Books. 2004.

Raeburn, J., and J. Jones. 1990. *The history of the international association of agricultural economists*. Brookfield: Gower Publishing Company.

Ricardo, D. 1821. *The principles of political economy and taxation*. London: M. Dent and Sons. 1911.

Roumasset, J. 1976. *Rice and risk: Decisionmaking among low income farmers*. Amsterdam: Elsevier.

Runge, C. 1981. Common property externalities: isolation, assurance and resource depletion in a traditional grazing context. *American Journal of Agricultural Economics* 63 (4): 595–606.

Runge, C. 1987. Induced agricultural innovation and environmental quality: the case of groundwater regulation. *Land Economics* 63: 249–258.

Runge, C., B. Senauer, P. Pardey, and M. Rosegrant. 2003. *Ending hunger in our lifetime: Food security and globalization*. Baltimore: Johns Hopkins University Press.

Ruttan, V. 1956. The contribution of technological progress to farm output, 1950–75. *The Review of Economics and Statistics* 38: 61–69.

Samuelson, P. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Schmitz, A., and D. Seckler. 1970. Mechanized agriculture and social welfare: the case of the tomato harvester. *American Journal of Agricultural Economics* 52: 569–577.

Schuh, G. 1974. The exchange rate and U.S. agriculture. *American Journal of Agricultural Economics* 56: 1–13.

Schuh, G. 1976. The new macroeconomics of agriculture. *American Journal of Agricultural Economics* 58: 802–811.

Schultz, T. 1953. *The economic organization of agriculture*. New York: McGraw Hill.

Schultz, T. 1964. *Transforming traditional agriculture*. New Haven: Yale University Press.

Schultz, T. 1971. *Investment in human capital*. New York: Free Press.

Scott, A. 1955. The fishery: The objectives of sole ownership. *Journal of Political Economy* 63: 116–124.
Segerson, K. 1988. Uncertainty and incentives for nonpoint pollution control. *Journal of Environmental Economics and Management* 15 (1): 87–98.
Sexton, R. 1990. Imperfect competition in agricultural markets and the role of Cooperatives: a spatial analysis. *American Journal of Agricultural Economics* 72: 709–720.
Sexton, R., and N. Lavoie. 2002. Food processing and distribution: An industrial organization approach. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, vol. 2B. Amsterdam: North-Holland.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Indianapolis: Liberty Classics. 1981.
Sumner, D., and S. Tangermann. 2002. International trade policy and negotiations. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, vol. 2B. Amsterdam: North-Holland.
Sunding, D., and D. Zilberman. 2002. The agricultural innovation process: Research and technology adoption in a changing agricultural sector. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, vol. 1A. Amsterdam: North- Holland.
Taylor, H. 1905. *An introduction to the study of agricultural economics*. New York: Macmillan.
Taylor, H. 1922. The history of the development of the Farm Economic Association. *Journal of Farm Economics* 4 (2): 92–100.
Taylor, H., and A. Taylor. 1952. *The story of agricultural economics in the United States, 1940–1932*. Ames: Iowa State University Press.
Timmer, C. 2002. Agriculture and economic development. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, vol. 2A. Amsterdam: North-Holland.
Tolley, H., Black, J. and Ezekiel, M. 1924. *Input as related to output in farm organization and cost of production studies*, Bulletin no. 1277. Washington, DC: USDA.
Tomek, W., and K. Robinson. 1977. Agricultural price analysis and outlook. In *A survey of agricultural economics literature*, ed. L. Martin, vol. 1. Minneapolis: University of Minnesota Press.
Veblen, T. 1900. The preconception of economic science, III. *Quarterly Journal of Economics* 14: 240–269.
Von Thünen, J. 1828. *Isolated state (English translation)*. Oxford: Oxford University Press. 1966.
Walker, J., R. Gardner, and E. Ostrom. 2000. Collective choice in the commons: experimental results on proposed allocation rules and votes. *Economic Journal* 110: 212–234.
Waugh, F. 1928. Quality factors influencing vegetable prices. *Journal of Farm Economics* 10 (2): 185–196.
Waugh, F. 1951. The minimum-cost dairy feed. *Journal of Farm Economics* 33: 299–305.
Working, H. 1922. *Factors determining the price of potatoes in St. Paul and Minneapolis,* Agricultural experiment station technical bulletin no. 10. University of Minnesota.
Working, H. 1925. The Statistical determination of demand curves. *Quarterly Journal of Economics* 39 : 503–543. August
Working, E. 1927. What do statistical 'demand curves' show? *Quarterly Journal of Economics* 41: 212–235.
Wright, B. 2002. Storage and price stabilization. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, vol. 1B. Amsterdam: North-Holland.

Agricultural Finance

Michael R. Carter

Abstract

Economic analysis of agricultural finance has traditionally focused on access to capital in the agricultural sector. Key concerns have included patterns of non-price rationing in agricultural credit markets, the institutions and contracts that provide credit to agricultural producers, the implications of the conditions of capital access on agricultural growth and rural income distribution, and the role of the public sector in agricultural credit markets. More recently, the analysis of agricultural finance has expanded beyond these credit-centred concerns to consider systemic approaches to rural finance that address risk and insurance, savings services and the provision of credit.

Keywords

Adverse selection; Agricultural banks; Agricultural finance; Agricultural markets in developing countries; Asymmetric information; Credit markets in developing countries; Credit rationing; Grameen Bank; Green Revolution; Microcredit; Moneylenders in developing countries

JEL Classifications

Q1

Several structural features of the agricultural sector make agricultural finance and financial

markets distinctive. First, the demand for agricultural finance is potentially high. Agricultural production processes are roundabout, with outputs and returns coming months or even years (in the case of vineyards and tree crops) after expenditures on productive inputs. The extreme riskiness of agriculture further increases the demand for credit or other contracts that share the risk of the production process.

A second distinguishing feature of agricultural finance is that the organization of agricultural production makes it difficult to supply with financial services. In a classic paper, John Brewster (Brewster 1950) noted that agricultural production differed from industrial production because of its spatial dispersion and its heavy dependence on inherently random inputs provided by nature. These features create what more contemporary economic analysis would call agency problems, meaning that it is difficult for an outsider to either monitor directly the quality of labour and management on a farm, or to infer *ex post* the qualities of those inputs from final agricultural output. As Brewster and others have remarked, the result is that agriculture tends to be organized in small-scale units, with much of the labour and management provided by the residual claimant to the production process (that is, it is rare to find large-scale 'factories in the field' except in special historical circumstances, as discussed by Binswanger et al. Barham et al. 1995).

Excess Demand for Financial Services

Agriculture thus stands as a sector with potentially high demand for financial services coming from relatively small-scale, spatially disperse, hard-to-monitor firms. In the contemporary low-income countries of Asia, Africa and Latin America, where the vast majority of farming households operate tiny holdings of an acre or two, between 5 and 15 per cent of producers have formal financial contracts (Braverman and Huppi 1991). Others are observed to borrow from a variety of informal sources, typically at nominal interest rates well in excess of those charged by formal financial institutions (Braverman and Guasch 1986).

While these observations are not by themselves sufficient to identify an excess demand for financial services in agriculture, they are consistent with it. Bolstering this interpretation is the fact that the characteristics of agriculture conform closely to the assumptions that underlie the formal economic theory of credit rationing. The seminal analysis of Stiglitz and Weiss (1981) assumes precisely the sorts of information costs and asymmetries that typify an agricultural sector comprising numerous, spatially disperse firms producing a highly random output. As extended by Carter (1988), this theoretical perspective suggests that adverse incentive and selection effects will prevent competitive formal lenders raising interest rates to market clearing levels (because higher rates result in lower expected profits for lenders as the borrowers still left in the market become increasingly less desirable as clients as interest rates increase). The result, according to this theory, is an agricultural credit market characterized by excess demand for formal credit and by a skewed allocation of (relatively cheap) formal credit toward larger farm units.

Some of this residual excess demand would be expected to spill over to locally based informal agents (moneylenders, input suppliers and processors). These lenders typically enjoy the twin advantages of cheaper information (because they are local) and the capacity to accept collaterals that could not be easily claimed by distant lenders (such as standing crops). Whether these agents are competitive suppliers of credit, or whether they enjoy spatial monopolies that grant them real market power, remains an open question (see, for example, Kochar 1997; Bell et al. 1997).

Implications of Excess Demand

While there is thus still debate about the degree of excess demand for financial services in agriculture, its implications are potentially large at two levels. First, excess demand for finance may result in slower agricultural technological change and growth. Again, examples from low-income countries make this point most easily. A study of new, input-intensive agricultural export products in

Central America found that annual working capital requirements per hectare exceeded the total annual incomes that farm families had been earning (Barham et al. 1995). The questionable ability of these families to self-finance investments of this magnitude, and to self-insure against the estimated 25 per cent failure rate of these activities, makes clear the economic costs of excess demand for agricultural finance. The deep and well-developed literature on the constrained adoption of inputintensive Green Revolution technologies ratifies this point.

In addition to its effects on the level and growth of agricultural incomes, excess demand for agricultural finance may also have impacts on income distribution within the rural economy. The theoretical analysis of Eswaran and Kotwal (1986) is especially instructive in this regard. Using a single-period general equilibrium model, they show that skewed access to capital, which leaves lower-wealth producers with excess demand for credit, will shift land access and income away from small-scale producer households, despite the intrinsic labour monitoring advantages enjoyed by these producers. The result is an agricultural economy that produces less, and distributes it less equally, than it would in a world of perfect financial markets. Eswaran and Kotwal go on to show that, under these conditions, an agricultural economy can become a prisoner of its own history. Economies that begin with relatively unequal wealth distributions tend to maintain them, while initially more egalitarian economies create more equal income distributions.

More recent theoretical analysis has used dynamic methods to extend the Eswaran and Kotwal analysis, asking whether the effects of excess demand for credit will be so long-lived and dramatic when credit-constrained and other agents have the option of building up their own sources of self-finance via savings over time. While not explicitly focused on agriculture, the analysis of Banerjee and Newman (1993) was an important demonstration that inadequate access to capital can fundamentally distort the occupational and production structure of an economy over the long term. Subsequent work has continued to build on this analytical tradition and has, among other things, shown that inadequate access to capital (in the presence of risk) can lead to a type of structural bifurcation in the agricultural economy. Initially wealthier producers move to a higher level of equilibrium well-being, while the initially poor become mired in a low-level poverty trap (see, for example, Dercon 1998; Mookherjee and Ray 2000; Zimmerman and Carter 2003).

Policy Debates

While much of this literature on the costs of inadequate access to capital in agriculture is relatively recent, the sense that agricultural financial markets are fundamentally imperfect has driven generations of policy interventions in both high and low-income nations. Historically, these interventions have included the direct provision of agricultural credit by public lenders, often at subsidized rates. For example, in the United States in 2002 more than 40 per cent of all farm debt to institutional lenders was held by two public entities, the Farm Credit System and the Farm Service Agency (USDA 2004). While still large, the public provision of agricultural credit in the United States has been trending downward for sometime, signalling the even larger role played by state credit in an earlier era when farms in the United States were smaller and more numerous.

In the low-income countries of Asia, Africa and Latin America, state agricultural banks and other mechanisms of public credit provision became a common feature of the agricultural landscape in the 1960s and 1970s. Interest rates were typically subsidized, and these interventionist policies were justified on the grounds that private provision of capital was either inadequate, priced at extortionate terms, or simply unavailable, especially for smaller farmers.

However, by the early 1980s, a coherent critique of these policies had emerged, arguing that state banks were financially unsustainable, crowded out private financial institutions, and did not even succeed in channelling credit to small-scale agricultural producers (see Adams et al. 1984). Under the pressure of structural adjustment and the broader move toward

economic liberalization, state agricultural banks began to disappear from the developing country landscape, and in Latin America, at least, were almost completely gone by the mid-1990s.

While commercial lending to agriculture continues to expand in the United States, the prediction by some that private institutional lenders would fill the gap left by public banks in Latin America and elsewhere in the developing world has been largely unfulfilled (Wenner et al. 2003). While in a few instances there has been renewed interest in public provision of agricultural finance, contemporary policy discussion largely focuses on three alternatives. The first is the provision of agricultural credit by non-financial businesses, such as input suppliers and commodity warehouses. The informational advantages of these informal lenders that permit them to monitor borrowers and lend where formal banks cannot has been more fully developed in recent theoretical literature (Conning 1999). As mentioned above, this sector remains enigmatic in terms of its efficiency and competitiveness. Nonetheless, there is increasing interest in the reform of collateral laws that might open the door to an expansion of lending by these businesses (Fleisig and de la Peña 2003). Others have argued that a general strengthening of legally weak landownership rights through systematic land titling programmes will induce greater entry into agricultural markets by private financial institutions (Feder and Akihiko 1999). However, evidence to date that land title bolsters formal credit supply to agricultural producers (especially small-scale producers) remains thin (Carter and Olinto 2003).

Micro-finance providers are a second alternative for the future provision of agricultural finance. Like informal lenders, micro-finance institutions can tap into cheap, locally available information about borrowers and their behaviour. They also utilize non-standard collateral assets, including group repayment guarantees in the case of micro-finance programmess that build on the Grameen Bank model of sequenced group loans. However, as Zeller and Meyer (2002) and others have discussed, the very localness of micro-finance institutions (which is the informational key to their ability to lend to small-scale, dispersed borrowers) can become a liability in weather-dependent agriculture where risks across borrowers are strongly correlated. Unlocking the potential for micro-finance lending to provide agricultural credit may thus require mechanisms to insure microfinance lenders, or their clients, against correlated weather risks. Pilot programmes to do just that are currently under development by the World Bank and others (Skees and Barnett 1999).

The third and final approach to the conundrum of agricultural finance is a more general systemic approach to developing rural (not necessarily agricultural) financial institutions. Motivated in part by the observation that farm families in both wealthy and developing nations derive much of their income from non-agricultural sources, this systemic approach advocates legal and institutional reforms designed to promote the expansion of full-service financial intermediaries in rural areas (Gonzalez-Vega 2003). Among these reforms are efforts to establish credit bureaus and other institutions that share borrowers' credit history across multiple lenders. Work such as that by Jappelli and Pagano (2002) suggests that the credit expansion effects of such institutions can be substantial. However, as with the other novel approaches described here, there is much yet to learn about whether these systemic approaches will suffice to improve the operation of financial markets in agriculture.

See Also

- ▶ Agricultural Markets in Developing Countries
- ▶ Credit Rationing
- ▶ Microcredit
- ▶ Moneylenders in Developing Countries

Bibliography

Adams, D., D. Graham, and J. von Pischke. 1984. *Undermining rural development with cheap credit*. Boulder: Westview Press.

Banerjee, A., and N. Newman. 1993. Occupational choice and the process of development. *Journal of Political Economy* 101: 274–298.

Barham, B., M. Carter, and W. Sigelko. 1995. Agro-export production and peasant land access, examining the dynamic between adoption and accumulation. *Journal of Development Economics* 46: 85–107.

Bell, C., T. Srinivasan, and C. Udry. 1997. Rationing, spillover and interlinking in credit markets: The case of rural Punjab. *Oxford Economic Papers* 49: 557–585.

Binswanger, H., K. Deininger, and G. Feder. 1995. Power, distortions, revolt and reform in agricultural and land relations. In *Handbook of development econo mics*, ed. J. Behrman and T. Srinivasan, Vol. IIIB. Amsterdam: North-Holland.

Braverman, A., and J. Guasch. 1986. Rural credit markets and institutions in developing countries: Lessons for policy analysis from practice and modern theory. *World Development* 14: 1253–1267.

Braverman, A., and M. Huppi. 1991. Improving rural finance in developing countries. *Finance and Development* 28: 42–44.

Brewster, J. 1950. The machine process in agriculture and industry. *Journal of Farm Economics* 32: 69–81.

Carter, M. 1988. Equilibrium credit rationing of small farm agriculture. *Journal of Development Economics* 28: 83–103.

Carter, M., and P. Olinto. 2003. Getting institutions right for whom? Credit constraints and the impact of property rights on the quantity and composition of investment. *American Journal of Agricultural Economics* 85: 173–186.

Conning, J. 1999. Outreach, sustainability and leverage in monitored and peermonitored lending. *Journal of Development Economics* 60: 51–77.

Dercon, S. 1998. Wealth, risk and activity choice, Cattle in Western Tanzania. *Journal of Development Economics* 55: 1–42.

Eswaran, M., and A. Kotwal. 1986. Access to capital and agrarian production organization. *Economic Journal* 96: 482–498.

Feder, G., and N. Akihiko. 1999. The benefits of land titling and registration: economic and social perspectives. *Land Policy Studies* 15: 25–43.

Fleisig, H. and de la Peña, N. 2003. Legal and regulatory requirements for effective rural finance markets. Paper presented at the Paving the Way Forward for Rural Finance Conference, Washington, DC, 2–3 June. Online. Available at http://www.basis.wisc.edu/rfc/documents/theme_legal.pdf. Accessed 21 Nov 2005.

Gonzalez-Vega, C. 2003. Deepening rural financial markets: Macroeconomic, policy and political dimensions. Paper presented at the Paving the Way Forward for Rural Finance Conference, Washington, DC, 2–3 June. Online. Available at http://www.basis.wisc.edu/rfc/documents/theme_macro.pdf. Accessed 21 Nov 2005.

Jappelli, T., and M. Pagano. 2002. Information sharing, lending and defaults: Crosscountry evidence. *Journal of Banking & Finance* 26: 2017–2045.

Kochar, A. 1997. An empirical investigation of rationing constraints in rural markets in India. *Journal of Development Economics* 53: 339–372.

Mookherjee, D., and D. Ray. 2000. Contractual structure and wealth accumulation. *American Economic Review* 92: 818–849.

Skees, J., and B. Barnett. 1999. Conceptual and practical considerations for sharing catastrophic/systemic risks. *Review of Agricultural Economics* 21: 424–441.

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

USDA (US Department of Agriculture). 2004. *Agriculture income and finance outlook/AIS–82*. Washington, DC: Economic Research Service, USDA.

Wenner, M., J. Alvarado, and F. Galarza. 2003. *Promising practices in rural finance, experiences from Latin America and the Caribbean*. Lima: Inter-American Development Bank.

Zeller, M., and R. Meyer. 2002. *The triangle of microfinance, outreach, financial sustainability and impact*. Baltimore: Johns Hopkins University Press.

Zimmerman, F., and M. Carter. 2003. Asset smoothing, consumption smoothing and the reproduction of inequality under risk and subsistence constraints. *Journal of Development Economics* 71: 233–260.

Agricultural Growth and Population Change

E. Boserup

The macroeconomic theory of the relationship between demographic and agricultural change was developed by Malthus and Ricardo in the early stage of demographic transition in Europe, and interest in classical theory was revived in the middle of this century, when economists became aware of the unfolding demographic transition in other parts of the world. Ricardo (1817) distinguished between two types of agricultural expansion in response to population growth. One is the extensive margin, the expansion into new land which he supposed would yield diminishing returns to labour and capital because the new land was presumed to be more distant or of poorer quality than the land already in use. The other type, the intensive margin, is more intensive cultivation of the existing fields, raising crop yields by such means as better fertilization, weeding, draining, and other land preparation. This also

was likely to yield diminishing returns to labour and capital. Therefore Ricardo assumed, with Malthus (1803), that population increase would sooner or later be arrested by a decline in real wages, increase of rents, and decline of per capita food consumption.

This theory takes no account of a third type of agricultural expansion in response to population growth: using the increasing labour force to crop the existing fields more frequently. This was in fact what was happening in England in Ricardo's time, when the European system of short fallow was being replaced by the system of annual cropping. Fallows are neither more distant nor of poorer quality than the cultivated fields, but if fallow periods are shortened or eliminated more labour and capital inputs are needed, both to prevent a decline of crop yields and to substitute for the decline in the amount of fodder for animals, which was previously obtained by the grazing of fallows. Therefore, this type of intensification is also likely to yield diminishing returns to labour and capital, but the additions to total output obtained by increasing the frequency of cropping are much larger than those obtainable by use of more labour and capital simply to raise crop yields. In fact, the Ricardian type of intensification is better viewed as a means not to raise crop yields, but more to prevent a decline of those yields as fallow is shortened or eliminated. When this third type of agricultural expansion by higher frequency of cropping is taken into account, elasticities of food supply in response to population growth are different from those assumed in classical theory.

The failure to take differences in frequency of cropping into account renders the classical theory unsuitable for the analysis of agricultural changes which accompany the demographic transition in developing countries in the second half of this century. Differences in population densities between developing countries are very large, and so are the related differences in frequency of cropping. The relevant classification for analysis of agricultural growth is not between new land and land which is sown and cropped each year, but the frequency at which a given piece of land is sown and cropped. Both in the past and today, we have a continuum of agricultural systems, ranging from the extreme case of land which is never used for crops, to the other extreme of land which is sown as soon as the previous crop is harvested. Increasing populations are provided with food and employment by gradual increase of the frequency of cropping.

In large, sparsely populated areas of Africa and Latin America, the local subsistence systems are pastoralism and long fallow systems of the same types as those used in most of Europe in the first millenium AD and earlier. In areas with extremely low population densities, twenty or more years of forest fallow alternate with one or two years of cropping, while four to six years of bush fallow alternate with several years of cropping in regions where population densities have become too high to permit the use of longer fallow periods. Methods of subsistence agriculture in developing countries with even higher population densities include short fallow systems (i.e. one or two crops followed by one or two years fallowing) or systems of annual cropping. In countries with very high population densities, including many Asian countries, some of the land is sown and cropped two or three times each year without any fallow periods.

If these differences in frequency of cropping are overlooked, or assumed to be adaptations to climatic or other permanent natural differences, the prospects for agricultural expansion in response to the growth of population and labour force look either more favourable, or more unfavourable, than they really are. In sparsely populated areas with long fallow systems, the areas which bear secondary forest or are used for grazing may be assumed incorrectly to be new land in the Ricardian sense, it being overlooked that they have the functions of recreating soil fertility or humidity, preventing erosion or suppressing troublesome weeds before the land is again used for crops. If neither the local cultivators nor their governments are aware of the risks of shortening fallow periods, and are not taking steps to avoid them, such shortening may damage the land and erosion, infertility or desertification may result. In such cases, the scope for accommodating increasing populations will prove to be less

than expected, and later repair of the damage will become costly, if possible at all. On the other hand, if land presently used as fallow in long fallow systems is assumed to be of inferior quality, in accordance with Ricardian theory, the large possibilities for accommodation of increasing populations by shifting from long fallow to shorter or no fallow, will be overlooked or underestimated.

Labour Supplies

When population growth accelerated in the developing countries in the middle of this century, economists applied Ricardo's distinction between expansion of cultivation to new land and attempts to raise crop yields by additional inputs of labour and capital. They therefore focused on the most densely populated countries in Asia, in which there was little new land. Since the possibilities for multicropping were not taken into account, it was assumed that the elasticity of food production in response to population growth would be very low in these countries, and that the acceleration of population growth would soon result in food shortages, high food prices, reduction of real wages, and steep increase of Ricardian rent.

Lewis (1954) suggested that in densely populated countries with little, if any uncultivated land, marginal returns to labour were likely to be zero or near to zero, and that a large part of the agricultural labour force was surplus labour, which could be transferred to non-agricultural employment without any diminution of agricultural output, even if there were no change in techniques. So Lewis recommended that rural-to-urban migration should be promoted, as a means of increasing marginal and average productivity in agriculture and of raising the share of the population employed in higher productivity occupations in urban areas. He confined his recommendation to densely populated countries, but many other economists made no distinction between densely and sparsely populated countries, assuming with Ricardo that uncultivated land must be of low quality so that a labour surplus would exist in all developing countries. The labour surplus theory contributed to create the bias in favour of industrial and urban development and the neglect of agriculture which has been a characteristic feature of government policy in many developing countries.

However, the labour surplus theory underestimates the demand for labour in agricultural systems with high frequency of cropping, based on labour intensive methods and use of primitive equipment. If population density in an area increases, fallow eliminated and multicropping introduced, then more and more labour-intensive methods must be used to preserve soil fertility, reduce weed growth and parasites, water the plants, grow fodder crops for animals, and protect the land. Some of the additional labour inputs are current operations, but others are labour investments. Before intensive cropping systems can be used, it may be necessary to terrace or level the land, build irrigation or drainage facilities, or fence the fields in order to control domestic animals. If these investments are made with human and animal muscle power, the necessary input of human labour is large. Even draught animals cannot reduce the work burden much, if fallows and other grazing land have been reduced so much that the cultivator must produce their fodder.

Part of the investments which are needed in order to increase the frequency of cropping are made by the cultivator with the same tools, animals and equipment that are used for current operations. Estimates of investments and savings in agricultural communities with increasing population are seriously low if they fail to include such labour investments. Due to the larger number of crops, the additional operations with each crop, and the labour investments, the demand for labour rises steeply when intensive land use is introduced. This contrasts with the assumptions of the labour surplus theory, which expects that the effect of population growth is always to add to the labour surplus.

When the theory of low supply elasticity and labour surplus in agriculture is combined with the theory of demographic transition, the prospects for densely populated countries with the majority of the population in agriculture look frightening. With the prospect of prolonged rapid growth of population (as forecast by the demographers) and

with the poor prospects for expansion of food production and agricultural employment (implied by the labour surplus theory), it seemed obvious that sufficient capital could not be forthcoming for the enormous expansion of non-agricultural employment and output that was needed. So because the possibilities for adapting food production to population were underestimated, many economists suggested that the best, or even the only means to avoid catastrophe was the promotion of rapid fertility decline by family planning. This in turn overlooked the links between the level of economic development and the motivations for restriction of family size.

The motivation for adopting an additional work load in periods of increasing population, and the means to shoulder it, are different as between agricultural subsistence economies and communities of commercial farmers. In the former, the need to produce enough food to feed a larger family may be sufficient motivation for adopting a new agricultural system which, at least for a time, raises labour input more steeply than output. The way to shoulder a larger work load is to increase the labour input of all family members. In some regions most of the agricultural work is done by men, and in other regions, by women; but when the work load becomes heavier, women become more involved in agricultural work in the former regions, and men more involved in the latter; in both, children and old people have more work to do. For all members of agricultural families, average work days become longer and days of leisure fewer. The whole year may become one long busy season in areas with widespread multicropping, labour intensive irrigation, and transplanting from seed beds.

For commercial producers, the motivation for intensification of agriculture emerges when population growth or increasing urban incomes increase the demand for food, and push food prices up until more frequent cropping becomes profitable, in spite of increasing costs of production or need for more capital investment. By this change in sectoral terms of trade, a part of the burden of rural population increase is passed on to the urban population. The increase of agricultural prices is by no means all an increase of Ricardian rent, but is in good part a compensation for increasing costs of production. If the increase of food prices is prevented by government intervention or by imports of cheap food, the intensification will not take place.

Moreover, in regions with commercial agriculture, work seasons become longer when crop-frequency increases in response to population growth. Therefore the decline of real wages per work hour is at least partially compensated for by more employment in the off-seasons, and by more employment opportunities for women and children in the families of agricultural workers. The discussion of low or zero marginal productivity in agriculture suffers from a neglect of the seasonal differences in employment and wages. Many off-season operations are in fact required in order to obtain higher crop-frequency through labour intensive methods alone, and so may well appear to be of very low productivity if viewed in isolation from their real function. Wages for these operations, or indeed off-seasons wages generally, may be very low; but the seasonal differences in wages are usually larger. Therefore, accumulation of debt in the off-seasons with repayment in the peak seasons is a frequent pattern of expenditure in labouring families.

Low off-season wages are an important incentive for intensification of the cropping pattern in commercial farms, since much of the additional labour with multicropping, irrigation, labour intensive crops and feeding of animals falls in these seasons. But, when the same land is cropped more frequently in response to population growth, the demand for labour in the peak seasons also rises steeply, perhaps more than the supply of labour. In many cases, a large share of the agricultural population combines subsistence production on small plots of owned or rented land with wage labour for commercial producers in the agricultural peak seasons, and this contributes to considerable flexibility in the labour market. If real wages decline, because population increase pushes food prices up, full time agricultural workers have no other choice than to reduce their leisure and that of their spouse and children, and offer to work for very low wages in the off-season periods. But workers, who have some

land to cultivate, may choose to limit their supply of wage labour, and instead cultivate their own land more intensively with family labour. Since they took wage labour mainly in the peak seasons, their limitation of the supply of wage labour may prevent a decline of, or cause an increase of, real wages in the peak seasons, and thus put a floor below the incomes of the full time workers.

The flexibility of the rural labour market is enhanced if not only labour but also land is hired in and out. A family that disposes of an increasing labour force may either do some work for other villagers, or rent some land from them, while a family that disposes of a reduced labour force may either hire some labour, or lease some land to others. With such a flexible system, prices for lease of land and wages will rapidly be adjusted to changes in labour supply. But the smooth adaptation of the system to population change will be hampered or prevented if, for political reasons, either hiring of labour or lease of land is made illegal, or changes in agricultural prices are prevented by government action.

Transport Cost and Urbanization

In Ricardian theory, marginal returns to labour and capital decline in response to population growth, partly because agricultural production is intensified, partly because it is expanded to inferior land, and partly because more distant land is taken into cultivation, thus increasing costs of transport. Thus, when population is increasing, producers have a choice between increasing costs of production, or increasing costs of transport between fields and consumers. However, there is a third possibility, which is to move the centre of consumption closer to land which is of similar quality to that which was used before the population became larger. Communities who use long fallow periods often move their habitations after long-term settlement in a forested area, and move to another area where the fertility of forest land has become high after a long period of non-use. Such movement of villages is likely to become more frequent, as population increases.

In other cases it is not the whole village which is moved, but an increasing number of villagers move their habitation to new lands, where they build isolated farmsteads or new hamlets. This may accommodate additional populations until all the space between the villages is filled up with habitations, and the choice in case of further population growth is between more frequent cropping, or use of inferior land, or long distance migration of part of the population.

The combination of shorter fallow periods and filling up of the space between the villages helps to create the conditions for emergence of small urban centres. Costs of transport are inversely related to the volume of transport, and roads, even primitive ones, are only economical, or feasible, with a relatively high volume of traffic. If fallow periods are very long, and distances between villages are large, there will be too few people in an area to handle both the production and transport which are necessary to supply a town with agricultural products. Urbanization and commercial agricultural production are only possible when population densities are relatively high, and fallow periods short. So when population in an area continues to increase a point may be reached when small market towns emerge, served by road and water transport, as happened in large parts of Europe in the beginning of this millenium.

With further growth of population it will again be necessary to choose between further intensification of agriculture at increasing costs, or moving the additional consumers (or some of them) to another location, where they can be supplied by less intensive agriculture, and with shorter distances of transport. So at this stage of development, new small market towns may emerge in between the old towns, or in peripheral areas together with agricultural settlement. In other words, instead of agricultural products moving over longer and longer distances, thus creating Ricardian rent in the neighbourhood of existing consumer centres, new centres of consumption may appear closer to the fields. In most of Europe, such a gradual spread of decentralized urbanization made it possible to delay the shift from short fallow agriculture to annual cropping to the late 18th or the 19th century. Areas with such a network of market towns have

better conditions for development of small-scale and middle-sized industrialization than sparsely populated areas with a scattered population of subsistence farmers.

The long-distance migration from Europe to North America in the 19th century can be viewed as a further step in this movement of European agricultural producers and consuming centres to a region with lower population density, less intensive agriculture, and much lower agricultural costs. The urban centres in America were supplied by extensive systems of short fallow agriculture at a time when production in Western Europe had shifted to much more intensive agriculture with annual cropping and fodder production.

Technology

From ancient times, growth of population and increase of urbanization have provided incentives to technological improvements in agriculture, either by transfer of technology from one region to another, or by inventions in response to urgent demand for increase of output, either of land, or labour, or both. Until the 19th century, technological change in agriculture was a change from primitive technology, that is, human labour with primitive tools, to intermediate technology, that is, human labour aided by better hand tools, animal-drawn equipment, and water power for flow irrigation. In the classical theory of agricultural growth, such changes are means to promote population growth and urbanization, but they are assumed to be fortuitous inventions, and are not viewed as technological changes induced by population growth and increasing urbanization.

In the course of the 19th century, the continuing increase of the demand for agricultural products, and the increasing competition of urban centres for agricultural labour, induced further technological change in European and North American agriculture. The technological innovations of the industrial revolution were used to accomplish a gradual shift from intermediate to high-level technologies, that is, human labour aided by mechanized power and other industrial inputs. The chemical and engineering industries contributed to raise productivity of both land, labour and transport of agricultural products, and scientific methods were introduced in agriculture as a means of raising yields of crops and livestock.

The existence of such high-level technologies improves the possibilities for rapid expansion of agricultural production in developing countries as well, but because in North America and Europe these technologies were used to reduce direct labour input in agriculture, those economists who believed in the labour surplus theory feared that they would further increase labour surplus. However, the idea of a general labour surplus in agriculture in developing countries had never been unanimously agreed, and under the influence of empirical studies of intensive agriculture in densely populated regions, Schultz (1964) suggested that labour was likely to be fully occupied even in very small holdings, when primitive technology was used. Therefore output and income in such holdings could only be increased by introduction of industrial and scientific inputs, and human capital investment of the types used in industrialized countries.

Although the proponents and the opponents of the labour surplus theory had different views concerning the relationship between the demand for and supply of labour, they agreed in suggesting a low supply elasticity of output in response to labour inputs, because they overlooked, or underestimated, the large effects on output and employment which can be obtained by using high-level technologies to increase the frequency of cropping. The availability of new varieties of quickly maturing seeds, of chemical fertilizers, and of mechanized equipment for pumping water and land improvements, permits the use of multicropping on a much larger scale, and in much drier and colder climates than was possible before these new types of inputs existed. The new high-level technologies have changed the constraints on the size of the world population from the single one of land area to those of energy supply and costs, and of capital investment.

The new inputs permit a much more flexible adaptation of agriculture to changes in population and real wages. Intensive agriculture is no longer linked to low real wages, and it is possible, by

changing the composition of inputs, to vary the rates of increase of employment and real wages for a given rate of increase of total output. By using a mixture of labour intensive and high-level techniques, adapted to the man–land ratio and the level of economic development, first Japan, and later many other densely populated countries, obtained rapid increases in both agricultural employment, output per worker, and total output. This 'Green Revolution' is an example of a technological change in agriculture induced by population change. The research which resulted in the development of these methods and inputs was undertaken and financed by national governments and international donors concerned about the effects of rapid population growth on the food situation in developing countries. Therefore, it focused mainly on improvement of agriculture in densely populated countries, where both governments and donors considered the problem to be most serious.

Agricultural producers who use high-level technologies are much more dependent upon the availability of good rural infrastructure than producers who use primitive or intermediate technologies. Transport and trade facilities are needed not only for the commercial surplus but also for the industrial inputs in agriculture; repair shops, electricity supply, technical schools, research stations, veterinary and extension services, are also needed. Therefore short-term supply elasticities differ between those regions which have and those which do not have the infrastructure needed for use of industrial and scientific inputs in agriculture. In the former, a rapid increase of output may be obtained by offering more attractive prices to the producers, while in the latter, increase of prices may have little effect on output, until the local infrastructure has been improved. Improvement of infrastructure may, on the other hand, be sufficient to obtain a change from subsistence production to commercial production, if it results in a major reduction in the difference between the prices paid to the local producers and those obtained in the consuming centres.

In densely populated regions with a network of small market towns, it is more feasible to introduce industrial and scientific inputs in agriculture, than in regions inhabited only by a scattered population of agricultural producers. Because per capita costs of infrastructure are lower in the first mentioned regions, they are more likely to have the necessary infrastructure, and if not, governments may be more willing to supply it. Thus sparsely populated regions are handicapped compared to densely populated ones, when high-level technologies are taken into use.

Tenure

Changes in output may also be prevented if the local tenure system is ill adapted to the new agricultural system. Land tenure is different in regions with different frequency of cropping. In regions with long fallow agriculture, individual producers have only usufruct rights in the land they use for cultivation, and the land, the pastures, and the forested land are all tribally owned. Before a plot is cleared for cultivation it is usually assigned by the local chief, and when large investments or other large works are needed the producers are organized by the chief as mutual work parties. If population increases and with it the demands for assignment of land, a stage may be reached when either the chief or the village community will demand a payment for such assignments, thereby changing the system of land tenure. Payments to the chief for assignment of land may turn him into a large scale landowner, and this payment may tip the balance and make more frequent cropping of land more economical than use of new plots, or settlement in new hamlets.

When frequency of cropping becomes sufficiently high that major permanent investments in land improvement are necessary, a change to private property in land may provide security of tenure to the cultivator, and make it possible for him to obtain credits. If at this stage no change of tenure is made by legal reform, a system of private property in land is likely to emerge by unlawful action and gradual change of custom; but in such cases the occupants, who have no legal rights to the land, may hesitate (or be unable) to make investments and land may remain unprotected against erosion and other damage.

In more densely populated areas, with more frequent cropping and need for large-scale irrigation and other land improvement, these investments may be organized by big landlords as labour service or by local authorities as wage labour, financed by local or general taxation. In order to change from a particular fallow system to another that is more intensive, it is likely that not only the ownership system in the cultivated plots but also that for uncultivated land must be changed, as must responsibility for infrastructure investment. Because of the links between the fallow system, the tenure system, and the responsibility for infrastructure investment, attempts to intensify the agricultural system by preservation (for political reasons) of the old tenure system and rural organization are likely to be unsuccessful, as are attempts to introduce new tenure systems that are unsuitable for the existing (or the desired, future) level of intensity and technology. Therefore, government policy is an important determinant of the agricultural response to population growth.

During fallow periods, the land is used for a variety of purposes: for gathering fuel and other wood, for hunting, for gathering of fertilizer, for grazing and browsing by domestic animals. Therefore, a change of the fallow system may create unintended damage to the environment unless substitutes are introduced for these commodities, or the pattern of consumption is changed. When hunting land becomes short, the right to hunt may be appropriated by the chiefs (or others), forcing the villagers to change their diet. When grazing land becomes short, enclosures may prevent the villagers (or some of them) from using it, or the village community may ration the right to pasture animals in the common grazing land and fallows, in order to prevent overgrazing and erosion, or desertification. These measures will impose a change of diet, and perhaps a change to fodder production in the fields.

Nutrition

Both production and consumption change from land-using to less land-using products when population increases and agriculture is intensified. There may be a shift from beef and mutton to pork and poultry, from animal to vegetable products, from cereals to rootcrops for human consumption, and from grazing to production of fodder for animals. Under conditions of commercial farming, the changes in consumption and production are induced by increasing differentials between the prices of land-saving and land-using products. If the process of population growth is accompanied by decline of real wages, the changes in consumption patterns for the poorest families may be large. This may result in protein deficiencies and malnutrition with spread of the disease–malnutrition syndrome; this causes high child mortality because disease prevents the child from eating and digesting food, and malnutrition reduces the resistance to disease.

The classical economists had suggested that continuing population growth would result in malnutrition, famine and disease, which would re-establish the balance between population and resources by increasing mortality. But they also envisaged the possibility of an alternative model, in which population growth was prevented by voluntary restraint on fertility. Malthus (1803) talked of moral restraint and Ricardo (1817) of the possibility that the workers would develop a taste for comforts and enjoyment, which would prevent a superabundant population. However, it was not ethical or psychological changes but the economic and social changes resulting from increasing industrialization and urbanization which induced a deceleration of rates of population growth, first in Europe and North America, and later in other parts of the world.

Government Policies

The deceleration of rates of population growth in Europe and North America coincided with a decline in the income elasticity of demand for food due to the increase in per capita incomes. As a result the rate of increase in the demand for food slowed down, just as the rate of increase of production accelerated due to the spread of high-level technologies and scientific methods in agriculture. If it had not been for government intervention in support of agriculture these changes

would have led to abandonment of production in marginal land, and use of less industrial inputs in the land that was kept in cultivation. But this process of adjustment was prevented by attempts to preserve the existing system of family farming. Large farms could utilize high-level technologies (especially mechanized inputs) better than smaller ones, but governments wanted to prevent the replacement of small or middle sized farms by larger capitalist farms, or company farming. Therefore, both Western Europe and North America gradually developed comprehensive systems of agricultural protection and subsidization of agriculture, agricultural research, and other rural infrastructure. In spite of this support a large proportion of the small farms disappeared and much marginal land went out of cultivation, while the support actually encouraged large farms, and farms in the most favoured regions, to expand their production; they increased their use of fertilizer and other inputs, and invested in expansion of capacity for vegetable and animal production. So supply still continued to outrun demand, and protection against imports and subsidies to exports still continued to increase, while the industrialized countries turned from being net importers to net exporters of more and more agricultural products.

In the discussions about labour surplus and low elasticity of agricultural production in non-industrialized countries, Nurkse (1953) had suggested that an increase in agricultural production could be obtained if the surplus population was employed in rural work projects. In the period until such a programme, in conjunction with industrialization and a deceleration of population growth, could re-establish the balance between demand for and supply of food, he recommended that temporary food imports (preferably as food aid) should be used to prevent food shortage. Because of the increasing costs of financing and disposing of the food surplus, Nurkse's suggestion of food aid was well received by Western governments, and transfer of food, as aid or subsidized exports, reached large dimensions.

Some governments in developing countries did use food aid and commercial imports of the food surpluses of the industrialized countries as stopgap measures, until their own promotion of rural infrastructure and other support to agriculture would make it possible for production to catch up with the rapidly increasing demand for food. But for many other governments the availability of cheap imports and gifts of food became a welcome help to avoid the use of their own resources to support agriculture and invest in rural infrastructure. Even in those developing countries with a large majority of the population occupied in agriculture, the share of government expenditure devoted to agriculture and related rural infrastructure is small, and within this small amount priority is usually given to development of non-food export crops, which often supply a large share of foreign exchange earnings. Exports of food crops are unattractive because of the surplus disposal of the industrialized countries, which exerts a downward pressure on world market prices. Therefore, both producers and governments in developing countries focus on the types of crops which do not compete with these subsidized exports. In regions in which the necessary infrastructure was available, employment and output of such export crops increased rapidly, not only in countries with abundant land resources but also in many densely populated countries, which shifted in part from food to non-food crops.

This general shift from food to non-food crops contributed to a downward pressure on export prices of the latter crops in the world market.

Food imports can have important short-term advantages for the importing country. Rapidly increasing urban areas can be supplied at low prices and without the need to use government resources to obtain expansion of domestic production. Moreover, counterpart funds from food aid can be used to finance general government expenditure, and in countries with high levies on export crops, government revenue increases when production is shifted from food to export crops. However, although there might be short-term advantages of food imports and food aid, the long-term cost of neglecting agricultural and rural development can be very high. The lack of transport facilities and local stocks, and the lack of irrigation in dry and semi-dry areas, may transform years of drought to years of famine. When governments do not invest in rural infrastructure and fail to provide the public services

which are necessary for the use of high technology inputs, the latter can be used only by large companies (who can themselves finance the necessary infrastructure) or in a few areas close to large cities.

Without cost reduction by improvement of the transport network and agricultural production, commercial food production may in many areas be unable to compete with imports. Commercial production will decline and subsistence producers will not become commercial producers. Instead, the most enterprising young villagers will emigrate in order to earn money incomes elsewhere. A larger and larger share of the rapidly increasing urban consumption must be imported, and food imports become a drug on which the importers become more and more dependent. The increasing dependency of many developing countries on food imports and food grants is often seen as a confirmation of the classical theory of inelastic food supply, and an argument for continuation of the policy of production subsidies and surplus disposal in America and Western Europe. Food imports are seen as gap fillers, bridging over increasing differences between food consumption and national food production in developing countries; but in many cases the gap is actually created by the food imports, because of their effect on local production and rural development.

Fertility

Contrary to the expectations prevalent in the middle of this century, government policy has proved to be a more important determinant of agricultural growth than the man–resource ratio, and the response to rapid population growth has often been better in densely populated countries than in sparsely populated ones with much better natural conditions for agricultural growth. The differences in agricultural growth rates and policies have in turn contributed to create differences in demographic trends, partly by their influence on industrial and urban development and partly by the effects on rural fertility, mortality and migration.

Because of their preoccupation with the man–land ratio, governments in densely populated countries not only devoted more attention and financial resources to agriculture than governments in sparsely populated countries, they also more often devoted attention and financial resources to policies aimed at reducing fertility. Moreover, tenure systems in densely populated countries usually provided less encouragement to large family size than tenure systems in sparsely populated countries.

In many densely populated countries with intensive agricultural systems, much of the rural population consists of small and middle-sized landowners, and such people are more likely to be motivated to a smaller family size than are landless labour and people with insecure tenure. They are less dependent upon help from adult children in emergencies and old age, because they can mortgage, lease, or sell land, or cultivate with hired labour. They may also have an interest in avoiding division of family property among too many heirs. If they live in areas where child labour is of little use in agriculture, they may have considerable economic interest in not having large families, and be responsive to advice and help from family planning services.

In sparsely populated regions with large landholdings, the rural population seldom has access to modern means of fertility control, and motivations for family restrictions are weak. A large share of the rural population tends to be landless or nearly landless workers, and if not they may be without security in land. So they are much more dependent upon help from adult children in emergencies and old age than are landowners, or tenants with secure tenure. If, moreover, their children work for wages in ranches, farms and plantations, the period until a child contributes more to family income than it costs is too short to provide sufficient economic motivation for family restriction.

People who use long fallow systems in regions with tribal tenure have even more motivation for large family size than landless workers. The size of the area they can dispose of for cultivation is directly related to the size of their family, and most of the work, at least with food production, is done by women and children. So a man can become rich by having several wives and large numbers of children working for him. Moreover, unless he has acquired other property a man's security in old age depends on his adult children and younger

wives, since he cannot mortgage or sell land in which he has only usufruct rights. Because of the differences in motivations for family size provided by individual and tribal tenure systems, the start of the fertility decline in regions with long fallow systems is likely to be linked to the time when population increase induces the replacement of the tribal tenure system by another system of tenure, and a decline is then more likely if it is replaced by small-scale land ownership than if it is replaced by large-scale farming.

In addition to the tenure system, changes in technological levels in agriculture and the availability of economic and social infrastructure may influence the timing of fertility decline in rural areas. The heavy reliance upon female and child labour in those densely populated areas in which agriculture is intensified by means of labour alone, may provide motivation for large families in spite of the shortage of land. Introduction of higher level technologies may then, in such cases, reduce a man's motivation to have a large family because it reduces the need for female and child labour. Use of intermediate and high-level technologies is nearly always reserved for adult men, while women and children do the operations for which primitive technologies are used. So when primitive technologies are replaced by higher level ones in more and more agricultural operations, men usually get more work to do and the economic contributions of their wives and children decline, thus reducing their economic interest in large family size. Moreover, in regions with little rural development high rates of child mortality may delay fertility decline, and the large-scale migration of youth from such areas may have a similar effect if parents can count on receiving remittances from emigrant offspring.

However, the relationship between rural development and fertility is complicated. Parents may want a large family for other than economic reasons, and increases in income due to rural development or to better prices for agricultural products make it easier for them to support a large family, thus preventing or delaying fertility decline. Other things being equal, fertility is positively related to income; but in developing societies most increases in income are caused and accompanied by technological, occupational and spatial changes that tend to encourage fertility decline, and the operation of these opposing effects may result in a relatively long time lag between rural modernization and fertility decline.

See Also

- ▶ Demographic Transition
- ▶ Hunting and Gathering Economies
- ▶ Labour Surplus Economies
- ▶ Malthus's Theory of Population
- ▶ Nutrition
- ▶ Peasants
- ▶ Thünen, Johann Heinrich von (1783–1850)

Bibliography

Boserup, E. 1965. *The conditions of agricultural growth*. London: Allen & Unwin.

Boserup, E. 1981. *Population and technological change*. Chicago: Chicago University Press.

Lewis, W.A. 1954. Economic development with unlimited supplies of labour. *Manchester School of Economic and Social Studies* 22(2): 139–191.

Malthus, T.R. 1803. *An essay on population*. London: J.M. Dent, 1958.

Nurkse, R. 1953. *Problems of capital formation in underdeveloped countries*. Oxford: Oxford University Press.

Ricardo, D. 1817. *The principles of political economy and taxation,* ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Schultz, T.W. 1964. *Transforming traditional agriculture*. New Haven: Yale University Press.

Schuttjer, W., and C. Stokes (eds.). 1984. *Rural development and human fertility.* New York: Macmillan.

Agricultural Markets in Developing Countries

Christopher B. Barrett and Emelly Mutambatsere

Abstract

The history of agricultural markets in developing countries reflects attempts to establish the appropriate government responses to the

inefficiencies created by incomplete institutional and physical infrastructure and imperfect competition. Government intervention in the 1960s and 1970s to resolve market failures gave way in the 1980s to market-oriented liberalization to 'get prices right' and, more recently, to 'get institutions right'. But market openness may accentuate the latent dualism of a modern, efficient marketing sector, accessible only to those with adequate scale and capital, alongside a traditional, inefficient marketing channel to which the poor are effectively restricted.

Keywords

Agricultural markets in developing countries; Agriculture and economic development; Border parity pricing; Commodity prices; Contract enforcement; Contract farming; Cooperatives; Corruption; Credit unions; Dual economies; Enforceability of contracts; First fundamental theorem of welfare economics; Foreign direct investment (FDI); General Agreement on Tariffs and Trade (GATT); Green Revolution; Imperfect competition; Imperfect information; Incomplete contracts; Infrastructure; Liberalization: of agriculture; Liquidity constraints; Market failure; Marketing boards; Monopsony; Non-tariff barriers; Oligopsony; Price controls; Price stabilization; Property rights; Public goods; Rule of law; Structural adjustment; Subsidies; Supermarkets; Tariffs; Transaction costs; World Trade Organization (WTO)

JEL Classifications

O1

Markets aggregate demand and supply across actors at different spatial and temporal scales. Well-functioning markets ensure that macro and sectoral policies change the incentives and constraints faced by micro-level decision makers. Macro policy commonly becomes ineffective without market transmission of the signals sent by central governments. Similarly, well-functioning markets underpin important opportunities at the micro level for welfare improvements that aggregate into sustainable macro-level growth. For example, without good access to distant markets that can absorb excess local supply, the adoption of more productive agricultural technologies typically leads to a drop in farm-gate product prices, erasing all or many of the gains to producers from technological change and thereby dampening incentives for farmers to adopt new technologies that can stimulate economic growth. Markets also play a fundamental role in managing risk associated with demand and supply shocks by facilitating adjustment in net export flows across space and in storage over time, thereby reducing the price variability faced by consumers and producers. Markets thus perform multiple valuable functions: distribution of inputs (such as fertilizer, seed) and outputs (such as crops, animal products) across space and time, transformation of raw commodities into value-added products, and transmission of information and risk. Per the first welfare theorem, competitive market equilibria help ensure an efficient allocation of resources so as to maximize aggregate welfare.

The micro-level realities of agricultural markets in much of the developing world, however, include poor communications and transport infrastructure, limited rule of law, and restricted access to commercial finance, all of which make markets function much less effectively than textbook models typically assume. A long-standing empirical literature documents considerable commodity price variability across space and seasons in developing countries, with various empirical tests of market integration suggesting significant and puzzling forgone arbitrage opportunities, significant entry and mobility barriers, and highly personalized exchange (Barrett 1997; Platteau 2000; Fackler and Goodwin 2001; Fafchamps 2004). Widespread inefficiencies result from incomplete or unclear property rights, imperfect contract monitoring and enforcement, high transactions costs, and binding liquidity constraints. Such failures often motivate government intervention in markets, although interventions have often done more harm than good, either by distorting incentives or by creating public sector market power. The history of agricultural markets in developing countries reflects evolving thinking

on the appropriate role for government in trying to address the inefficiencies created by incomplete institutional and physical infrastructure and imperfect competition. The emphasis in the 1960s and 1970s on government intervention to resolve market failures gave way in the 1980s to market-oriented liberalization to 'get prices right' and, more recently, to a focus on 'getting institutions right'.

Past Approaches

Agricultural marketing of most major export and food commodities and of modern inputs – such as fertilizer, machinery and hybrid seed – was historically highly regulated by developing country governments into the 1980s, via input price controls and subsidies, oligopolistic input markets, monopsonistic produce marketing boards, pan-seasonal and pan-territorial administrative commodity pricing, oligopolistic processing industries, and fixed wholesale and retail prices. Commodity prices were generally set below market levels, implicitly taxing producers while subsidizing consumers. Marketing channels were typically very inefficient, with centralized storage and processing facilities and government-imposed grades and standards for product quality, although these were not always and everywhere enforced. Sometimes these inefficient systems provided satisfactory coordination of marketing channels, but that was by no means universal. Heavy government presence, especially pan-seasonal and pan-territorial producer pricing, and fixed retail pricing systems and bans on private commerce effectively eliminated most incentives for private arbitrage or investment in fixed capital by marketing intermediaries. Meanwhile, management by government fiat too often facilitated corruption, which often had a devastating long-run impact on economic governance.

In addition to state-run marketing boards, producer marketing cooperatives were prevalent in developing countries at all levels of the marketing chain, ranging from credit unions through farmer cooperatives to wholesale-level cooperatives. Credit unions commonly accumulated funds for input purchase or served as intermediaries for government-subsidized credit programmes. Farmer marketing cooperatives typically facilitated bulk input procurement, price negotiation, and sharing of transportation costs. Wholesale cooperatives mainly assembled bulk commodity lots for sale into government processing and distribution channels. Cooperatives have often worked well in specialized production areas distant from major markets, and with homogenous production of not-so-perishable commodities such as coffee. However, due to high administrative and coordination costs, free-rider problems and political interference, cooperative systems have not lived up to expectations in most developing countries, and many have collapsed.

In contrast to the major export and domestic staple food crops, smaller-scale food commodities for domestic consumption, such as indigenous fruits and vegetables, have almost always operated on a free market basis, with little history of state intervention or price regulation. These markets are characterized by many cash, spot market transfers of product between intermediaries en route from producer to consumer, many small, non-specialized and unorganized buyers and sellers, few if any grades or standards, one-on-one (dyadic) price negotiations, poor market information systems, and mostly informal contracts, largely enforced through social networks (Fafchamps 2004). Such marketing channels depend disproportionately on rural periodic markets prevalent in most of the developing world, arguably the closest one ever gets to a true 'free market': free of government regulation, subsidies and taxes, and lacking public goods such as physical infrastructure, contract law, public market price information systems, or codified product grades and standards. Indeed, they have been termed the 'flea market economy' by Fafchamps and Minten (2001).

The Emerging Problems of State Agricultural Market Control

Given the inherent variability of agricultural production and the significance of agriculture in economic activity and general well-being in

developing countries, price stabilization policies were long considered necessary for economic stability. However, a number of problems emerged. First, the fixing of commodity prices below market levels inevitably created a disincentive for agricultural producers. By the late 1970s, low producer prices had led to the stagnation of production and exports and to increased parallel market activity, including cross-border smuggling, in many developing countries, especially in those areas of Africa and Central America that were largely bypassed by the Green Revolution.

The second major problem was the fiscal and political sustainability of government agricultural market interventions. The inefficiencies of parastatal marketing boards, along with the repression of private market intermediation, led to unreliable supplies of consumer goods for politically important urban populations. Moreover, those inefficiencies, combined with the numerous subsidies and frequent corruption within government-controlled marketing channels, became too costly for central governments, which faced massive pressure from international donors in the 1980s and 1990s to trim expenditures and to eliminate price controls (Timmer 1986).

Economic Liberalization: Market Relaxation and State Compression

Market-oriented agricultural policy reforms were a centrepiece of economic liberalization in developing countries in the 1980s and 1990s, commonly within the context of broader structural adjustment programmes designed to restore fiscal and current account balance, to reduce or eliminate price distortions, and to facilitate efficient price transmission so as to stimulate investment and production. The new focus was on re-establishing a close correspondence between local and world market prices, so-called border parity pricing. The withdrawal of the state from agricultural market intermediation, specifically price discovery, was seen as a necessary condition in getting prices right, itself a necessary condition for improving market efficiency and stimulating investment and productivity growth (Timmer 1986).

The market-oriented reforms typically implemented by developing country governments included, on the input side, the liberalization of land and labour markets, decontrol and de-licensing of input production, supply and distribution, removal of input subsidies and price controls, closure of loss-making credit schemes, liberalization of credit markets, and reform of agricultural extension. On the output markets side, reforms included commodity price liberalization, the removal of parastatal monopoly power and commodity movement restrictions, and reduction in tariffs and quotas on imports.

The net result of these reforms typically turned on the balance between the pro-competitive effects of reduced government interference in marketing operations – what Lipton (1993) termed 'market relaxation' – and the anti-competitive effects of reduction of public goods and services that underpin private market transactions – what Lipton (1993) termed 'state compression'. Since the two phenomena were typically inextricable in agricultural liberalization initiatives, experiences varied markedly.

The empirical evidence suggests that commodity prices generally increased after market reforms, often stimulating an increase in production, especially of export crops. These price increases also facilitated the emergence of supermarket chains, export-oriented outgrower schemes and export processing zones, and a generalized stimulus to agro-industrialization in developing countries (Reardon and Barrett 2000; Sahn et al. 1997). Increased investment in the downstream marketing channel has transformed the orientation of many agricultural markets from raw commodity towards processed product markets, and with this increased investment came increased competition. In countries such as Chile, India and South Africa, private firms now play a leading role in development of improved seed varieties, producing and distributing inputs, post-harvest processing and modern retailing through supermarkets and restaurant chains (Reardon et al. 2003; Reardon and Timmer 2005). Both formal and informal traders entered agricultural commodity marketing channels as government controls fell away, from rural periodic markets all the way through urban retail markets.

However, market entry has tended to be limited to certain marketing niches not protected by capital, information or relationship barriers, with substantial bottlenecks in other areas such as inter-seasonal storage and motorized transportation. Neither widespread entry into market intermediation activities nor workably competitive markets emerged everywhere, let alone quickly. For example, because long-haul motorized transportation in rural markets tends to involve considerable sunk costs and some economies of scale due to poor road conditions and high vehicle maintenance costs, entry into this sector of the markets has often been limited after the removal of legal and policy barriers to entry (Barrett 1997). Meanwhile, the end of pan-seasonal and pan-territorial administrative pricing has brought increased price risk, with consequences for investment incentives facing both producers and market intermediaries (Barrett and Carter 1999).

The elimination of input subsidies and removal of government monopsony power in crop marketing has also often led to reduced access to input financing and increased input prices. The withdrawal of parastatals from core input marketing activities created a void that the private sector often failed to fill due to underdeveloped physical communications, power and transport infrastructure, credit constraints and continued bureaucratic impediments that increased transactions costs for input suppliers. In addition, periodic state and donor-funded input programmes have often reduced profitability and frustrated private investments. Input credit schemes by processors have been used in the post-reform period in an attempt to overcome the low input use resulting from these access problems, for example in the cotton sectors of Mali and Uganda and horticultural export sectors of Kenya and Zimbabwe.

Although the level of reform implementation differed from country to country, in many cases reform was only partially implemented and policy reversals were common (Jayne and Jones 1997; Kherallah et al. 2002). In important food and export markets, liberalization efforts have been prolonged and incomplete, reflecting the difficulty in relinquishing government control in the face of uncertainty and political pressures to intervene in order to resolve perceived inequities or inefficiencies in market performance. For example, parastatals remain active in the West African cotton sector, the southern African maize sector has not been fully liberalized, and in Indonesia BULOG continues to operate amid private marketing companies. The ebb and flow of market-oriented reforms and the frequency with which governments have engaged in policy reversals has made it terribly difficult to tease out clear patterns in the impact of liberalization measures on the performance of agricultural markets in developing countries.

Post-structural Adjustment Market Reforms

As the weaknesses of reformed agricultural markets in developing countries became evident, development agencies' and governments' focus began to shift from merely 'getting prices right' to 'getting institutions right' so as to address market failures arising from imperfect information, contract enforcement and property rights, and insufficient provision of public goods. Such reforms have used non-price measures in an attempt to develop the public and private institutions necessary for efficient market operations and to reduce transactions costs and business risk.

The post-structural adjustment era has also coincided with international market deregulation through the GATT and its successor, the WTO. Bilateral, regional and global trade agreements have reduced tariff and non-tariff barriers to cross-border flows of raw and processed agricultural commodities, and increased the openness of financial markets, leading to increased capital flow into developing countries, especially in the form of foreign direct investment (FDI). Where structural adjustment reforms had substantially reduced state control over input and output markets, trade and FDI liberalization has paved the way for major investment in post-harvest processing and retailing in developing countries since the 1990s. This 'new' capital investment differs from the structural adjustment era reforms in that whereas the focus previously was

upstream, in the input, production, and wholesale sectors, more recent emphasis, especially in private investment, has tended to be downstream, in food processing, retail and restaurant markets. The exceptionally rapid diffusion of supermarkets in developing countries, in particular, has also been driven by improved coordination and communication technologies in addition to increased urbanization, lower prices of processed goods, increased per capita incomes in developing countries, as well as saturation and intense competition in foreign firms' home markets (Reardon and Barrett 2000; Reardon et al. 2003). In Latin America, for example, supermarkets currently account for 50–60 per cent of national food retail sale, compared with only 10–20 per cent in the 1980s (Reardon et al. 2003; Reardon and Timmer 2005).

The rise of supermarket and restaurant chains has changed the fundamental structure and operations of agricultural markets significantly, directing far more market power downstream, often to chains wholly or partly owned by multinational corporations. Commodity procurement by retailers has become more centralized, with consolidated buying points at a regional, even global, level. It is not uncommon for a major supermarket chain located in three different countries to consolidate its procurement in a few large growers in just one of those countries. Global food chains have also established regional procurement nodes – for example, Walmart throughout Asia and Latin America – and in-country commodity procurement for regional firms such as the China Resource Enterprise has been centralized from individual store level to provincial systems (Reardon et al. 2003). These structural shifts have increased contract farming and outgrower schemes between agro-industrial firms and farmers in developing countries, and production of non-staple foods has increased.

Increased foreign investment in agricultural markets in developing countries, however, has produced conflicting results. Increased industrialization of agricultural markets has fostered improved market efficiency and competitiveness, integration of formerly fragmented markets, product diversification through differentiation, and value addition and technology transfer. However, the rapid pace of structural change, with some developing countries accomplishing in a few years what developed countries accomplished over decades, has left limited room for adjustment by smaller, less well-informed and poorly capitalized market actors to new ways of doing business. There is thus growing concern that market openness may lead to the replacement of traditional processors by oligopsonistic multinationals, accentuating the latent dualism of a modern, efficient marketing sector accessible only to those with adequate scale and capital, alongside a traditional, inefficient marketing channel to which the poor are effectively restricted. The tendency towards selection of a few medium-to large-scale firms or producers capable of delivering consistent quality product at large volumes has toughened competition for structurally inefficient producers, and seems to have led to some crowding out of smaller producers (Reardon and Timmer 2005). Local informal wholesalers and retailers have found themselves having to compete with bigger firms, both for the more efficient producers offering consistent product quality and throughput volumes, and for consumers seeking more services. The emergence of big, concentrated downstream private marketing intermediaries could also potentially lead, once again, to non-competitive agricultural marketing channels, effectively replacing government with private market power.

Increased contract farming, while offering significant potential for smaller growers in the form of guaranteed markets and prices for their produce often coupled with input credit and extension service, has evidently also reduced farmer bargaining power in negotiating contract conditions. These negotiations now take place bilaterally, between individual farmers and the large contracting firm, rather than via collective bargaining by farmer associations with government parastatals.

Conclusion

Agricultural markets play a crucial role in the process of economic development. Yet, by virtue of the spatial dispersion of producers and consumers, the temporal lags between input

application and harvest, the variable perishability and storability of commodities, and the political sensitivity of basic food staples, agricultural markets are prone to high transactions costs, significant risks and frequent government interference. The relative power of developing country governments and private domestic or multinational firms in agricultural markets has varied over time. But the fundamental functions of input and output distribution, post-harvest processing and storage, as well as the persistent challenges of liquidity constraints, contract enforcement and imperfect information, have characterized agricultural markets in developing countries under all forms of organization.

See Also

- ▶ Agriculture and Economic Development
- ▶ Development Economics
- ▶ Dual Economies
- ▶ Foreign Direct Investment
- ▶ Marketing Boards
- ▶ Spatial Market Integration

Bibliography

Barrett, C. 1997. Food marketing liberalization and trader entry: Evidence from Madagascar. *World Development* 25: 763–777.

Barrett, C., and M. Carter. 1999. Microeconomically coherent agricultural policy reform in Africa. In *African economies in transition, volume 2: The reform experiences*, ed. J. Paulson. London: Macmillan.

Fafchamps, M., and B. Minten. 2001. Property rights in a flea market economy. *Economic Development and Cultural Change* 49: 229–268.

Fafchamps, M. 2004. *Market institutions in Sub-Saharan Africa*. Cambridge, MA: MIT Press.

Fackler, P., and B. Goodwin. 2001. Spatial price analysis. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, Vol. 1B. Amsterdam: Elsevier.

Jayne, T., and S. Jones. 1997. Food marketing and pricing policy in eastern and southern Africa: A survey. *World Development* 25: 1505–1527.

Kherallah, M., C. Delgado, E. Gabre-Madhin, N. Minot, and M. Johnson. 2002. *Reforming agricultural markets in Africa*. Washington, DC: International Food Policy Research Institute.

Lipton, M. 1993. Market relaxation and agricultural development. In *States and markets: Neo-liberalism and the development policy debate*, ed. C. Colclough and J. Manor. Oxford: Oxford University Press.

Platteau, J.-P. 2000. *Institutions, social norms, and economic development*. London: Harwood Academic Publishers.

Reardon, T., and C. Barrett. 2000. Agroindustrialization, globalization, and international development: An overview of issues, patterns, and determinants. *Agricultural Economics* 23(3): 195–205.

Reardon, T., and P. Timmer. 2005. Transformation of markets for agricultural output in developing countries since 1950: How has thinking changed? In *Handbook of agricultural economics*, ed. R. Evenson, P. Pingali, and T. Schultz, Vol. 3A. Amsterdam: Elsevier.

Reardon, T., P. Timmer, C. Barrett, and J. Berdegue. 2003. The rise of supermarket chains in Africa, Asia and Latin America. *American Journal of Agricultural Economics* 85: 1140–1146.

Sahn, D., P. Dorosh, and S. Younger. 1997. *Structural adjustment reconsidered*. Cambridge: Cambridge University Press.

Timmer, P.C. 1986. *Getting prices right – The scope and limits of agricultural price policy*. Ithaca: Cornell University Press.

Agricultural Research

Robert E. Evenson

Abstract

This article reviews contributions made by agricultural research programmes in historical context. Before 1850, when the agricultural experiment station (AES) model was developed, most crop and livestock improvement was due to farmer selection of seeds and livestock breeding. By 1875, a number of plant breeding programmes were in place. Developed countries achieved a green revolution in the first half of the 20th century, developing countries in the second half. A number of countries are now benefiting from the gene revolution. An assessment of social returns to public spending on agricultural research shows these returns to be high.

Keywords

Agricultural research; Ehrlich, P.; Genetically modified (GM) crops; Green revolution; Green revolution modern varieties (GRMVs); Griliches, Z.; Internal rate of return (IRR); International agricultural research centers (IARCs); Mendel, G.; National agricultural research system (NARS); Patents; Precautionary principle; Recombinant DNA (rDNA) gene revolution; Returns to research

JEL Classifications

Q1

Before 1850

The earliest form of agricultural research was agricultural invention. The patent systems in Europe date back to the Statute of Monopolies in 1623 in England. During the 18th century, England and France further developed their patent systems. Article 1, Section B of the US Constitution, drawn up in 1787, states that 'Congress shall have the power to promote the progress of science and useful arts, by securing for limited times for authors and inventors the exclusive right to their respective writings and discoveries.' The first Patent Act in the United States was enacted in 1790. Many of the earliest inventions, including Eli Whitney's cotton gin, were agricultural inventions.

Prior to the development of the modern agricultural experiment station in 1843, the 'botanic garden' served as the chief research vehicle for plants. Botanic gardens were established in many countries, preserving and further classifying plants and trees in the tradition of Linnaeus. (Today there are 1,500 botanical gardens worldwide. Of these, 698 have germplasm collections for the conservation of ornamental species, indigenous crop relatives and medicinal and forest species, and 119 conserve germplasm of cultivated species, including landraces – that is, distinct types – and wild food plants.)

Both plant and animal improvement prior to the modern experiment station was achieved by farmers themselves. Prior to the 18th century, farmers selected seed from each crop to improve the productivity of crop species. (There are approximately 300,000 species of higher plants, that is, flowering and cone-bearing plants. Of these, 270,000 have been identified and described. About 30,000 species are edible and about 7,000 have been cultivated or collected by humans for food; 120 species are important cultivated crops, but 90 per cent of the world's caloric intake is provided by only 30 species.)

As populations moved to new locations and production conditions, they created new landraces in each cultivated species. As new landraces were created, three distinct classes were identified. Landraces created in the centre of origin of cultivation were the first class. For rice, as many as three or four centres of origin (that is, locations of first cultivation) for the two cultivated species *Oryza sativa* and *Oryza glaberrima* have been identified. The second class includes landraces created in centers of diffusion (that is, locations where populations diffused the crop). The third class comprised landraces created in the New World countries in the Americas and Oceania.

These landraces were later collected and, along with mutants and uncultivated species in the genus, they constitute the genetic resources used in modern plant breeding programmes based on conventional methods of crossing parental plants. Table 1 summarizes contemporary *ex situ* genebank collections.

Animal improvement actually pre-dates crop improvement. It, too, was achieved by farmers and herdsmen. Most of the breeds of cattle, pigs, poultry, horses, sheep, and so forth were developed in the 16th through 18th centuries. Most were developed in Europe. Work animals, including oxen, horses and water buffalo, were particularly important in agriculture prior to the 20th century, when tractors became the dominant source of power in many countries. Work animals, including the powerful workhorses, important to cultivation, are sensitive to climatic conditions. Animal breeds used in Asia range from the powerful bullocks in North India, weighing more than a ton, to much smaller cattle in the Himalayan mountains.

Agricultural Research, Table 1 Genebank collections (ex *situ)*

Crops	Estimated numbers of landraces (000's)	Major collections (number)	Genebank accessions (000's)	Percent in genebanks
Cereals				
Wheat	150	36	844	95
Rice	130	20	420	90
Maize	65	22	277	90
Sorghum	45	19	169	80
Millets	30	18	90	80
Legumes				
Beans	n/a	15	268	50
Soybeans	30	23	174	60
Lentils	n/a	5	26	n/a
Groundnuts	15	16	81	n/a
Root crops				
Cassava	n/a	5	28	35
Potato	30	16	31	95
Sweet potato	5	7	52	50
Other				
Sugar cane	20	20	20	70

Source: FAO (1998)

1850–1900

Agricultural research programmes were changed dramatically with the development of the agricultural experiment station. It is generally accepted that the first truly scientific experiment stations were located in the UK, in the Rothamsted Experiment Station, established in 1843, and in Saxony, where several experiment stations were established in the 1850s.

With the experiment station and its formal structure of experiments with 'treatments' and 'controls', agricultural research became scientific, and by 1900 agricultural science was established as a mature applied science. The application of statistical methods to experiments furthered this development. R. A. Fisher, the statistician at the Rothamsted Experiment Station in the UK from 1919 to 1933, is credited with numerous methodological developments, many of them relevant to modern-day econometrics. Early experiments focused on agricultural chemistry, including the application of chemical fertilizers and related soil amendments. By 1875 or so, formal plant breeding programmes were beginning to be established. It is often thought that formal plant breeding did not take place until after the 'rediscovery' of Gregor Mendel's work, first published in 1856, in 1900. But that is not the case: breeding programmes in sugar cane, wheat and many other crops were established before 1900. Sugar cane breeders in Java and Barbados simultaneously discovered techniques to induce flowering in sugar cane plants in 1878, and by 1900 the 'noble' canes from their breeding programmes were beginning to transform sugar cane production in several countries.

In the United States, the Hatch Act of 1887 provided funds for experiment stations in every state. Most state experiment stations recognized the synergistic relationship between research and graduate teaching, and formally linked experiment stations with land grant college programmes. It is widely thought that legislation such as the Hatch Act reflected exceptional wisdom on the part of legislators. This was not the case. Prior to the Hatch Act, many states had considerable experience with experiment stations. This was also true for the Land Grant College Act – the Morill Act – in 1862. Some 20 states had established colleges of

agriculture prior to 1862. As these programmes matured, veterinary medicine colleges were established in land grant colleges. By 1900, sufficient experimental data were available from state agricultural experiment stations to answer many questions of importance to farmers in the US.

1900–1940

The period 1900 to 1940 was a one of extraordinary achievements by agricultural experiment stations. Plant breeding gains were achieved in most crops planted in temperate zone countries (in effect, temperate-zone developed countries realized a green revolution in this period). Plant breeding gains in sugar cane, coffee, tea and spices (the Mother Country crops) were also achieved in tropical regions. Brazil and Argentina in Latin America realized major gains (Brazil became the world's major producer of coffee and sugar; Argentina the major exporter of beef).

Two major scientific developments in plant breeding were achieved during this period. The first was the development of techniques to produce hybrid crop varieties to take advantage of the 'heterosis' effect in crops. The early development of hybrid techniques took place at Harvard and Yale Universities, but the major achievement was made by Donald Jones at the Connecticut Agricultural Experiment Station in New Haven. Jones developed the 'double cross' method for seed production. Hybrid seed production requires 'selfing' or 'inbreeding' for several generations. Prior to Jones, a single cross was made between two inbred lines to produce hybrid seed; the seed cannot be saved by farmers because the heterosis effect is present only in the hybrid generation. Jones used four inbred lines in a double-cross to produce seed more efficiently. Since Connecticut is not a major corn production state, it was several years before hybrid corn was available to farmers in Iowa. Henry A. Wallace, later a vice-president of the US, was an early leader in developing private industry production of hybrid corn. He established the Pioneer Hybrid Seed Company in 1926.

Zvi Griliches (1957) analysed the adoption of hybrid corn by farmers in different US states. Farmers in Alabama had access to hybrid corn varieties 20 years after farmers in Iowa. This was not because hybrids suited to Iowa farmers were not exhaustively evaluated in Alabama. Alabama farmers did not have hybrid varieties until seed companies established breeding programmes in Alabama to develop varieties suited to Alabama production conditions. Corn has a high degree of photo-period sensitivity. Varieties suited to Alabama were also varieties with longer growing seasons. This same principle applies to the green revolution (see below). No country without a functioning plant breeding programme has realized a green revolution.

The second scientific development was another form of hybridization, interspecific hybridization or 'wide crossing'. Until the gene revolution, based on 'recombinant DNA' techniques, all plant breeding entailed a 'sexual' cross between two 'parent' cultivars (this continues to be the case for achieving continuous plant improvement). Inter-specific hybridization entails a sexual cross between different species, usually members of the same genus. This was first achieved in sugar cane in 1919 when breeders achieved crosses between *Saccharum officianaram,* the cultivated species, and *Saccharum spontaneum,* an ornamental species of sugar cane. Later a third species, *Saccharum barberie,* was added.

By the 1980s, inter-specific hybridization techniques (chiefly embryo rescue techniques) had been developed for most crop species. With these techniques, sexual crosses have been achieved between cultivated species and most or all uncultivated species in the same genus for all important crop species.

During 1900–40, developed country agriculture (and some developing country agriculture) was also being affected by the development of farm machinery and tractor power. Stationary tractors and steam engines were developed before 1900. After 1900 the row crop tractor was developed along with improved harvesting and planting machinery. By the 1930s these developments were changing the structure (farm size, off-farm work) of US agriculture. These developments were produced largely by private sector firms in the farm machinery and farm chemical industries.

Patent incentives existed for mechanical, electrical and chemical inventions in this period. They were not developed for genetic inventions until after 1980.

1940–1965

At the end of the Second World War, agricultural research experienced a renaissance in developed countries. This was at least in part because of synergism between public sector agricultural research and private sector R&D in the farm machinery and farm chemical industries. By 1965 supermarkets had crowded out the 'mom and pop' grocery stores in most US cities. Poultry production was effectively industrialized by 1965 as confined housing units became the norm. Dairy production was subject to scale economies, and herd size was increasing. Feed management had improved greatly. The widespread use of United States Department of Agriculture grades and standards for livestock was transforming the meat packing industry. By 1965, in all OECD countries total factor productivity growth was faster in the agricultural sector than in the rest of the economy, and this continues to be the case today.

In developing economies, a sense of alarm had been created by the growing recognition that developing countries were in for a population explosion. With improvements in public health measures, death rates, particularly among children, began to decline and life expectancy began to increase. With even modest delays until the birth rate declined, this meant rapid increases in population. The alarm in question centred on food security. Many alarmists of the 1950s, notably Paul Ehrlich (1968), concluded that food production growth could not keep pace with population growth.

The international community (including the World Bank, regional banks, foundations and bilateral aid organizations) responded by developing a system of international agricultural research centers (IARCs). The first two IARCs were the International Rice Research Institute (IRRI) in the Philippines and the International Wheat and Maize Improvement Center (CIMMYT) in Mexico. These two centres were credited with creating a 'green revolution' based on high-yielding varieties of rice and wheat introduced to farmers in 1965. Other IARCs, however, contributed to green revolutions in all major food crops.

The Green Revolution: 1965–2004

The period 1965–2004 was truly extraordinary for agriculture. In 1991 the Soviet Union collapsed, leaving the former Soviet republics in severe recession. This included the agricultural sector. Most, but not all, developing countries experienced a green revolution during this period.

Table 2 summarizes the production of green revolution modern varieties (GRMVs) by five-year period. These data show that the production of GRMVs is increasing over time. Thirty-six per cent of all GRMVs were crossed in an IARC programme. Twenty-two per cent of GRMVs crossed in national agricultural research system (NARS) programmes utilized an IARC-crossed parent or other ancestors. Non-government organizations (NGOs) did not produce GRMVs. None were crossed in developed country programmes and transferred to developing countries. Private sector firms did produce hybrid maize, sorghum and millet varieties (five per cent of GRMVs) but only after improved open-pollinated varieties (OPVs) had been produced by IARC programmes. GRMVs were produced in public sector IARC programs and in NARS programmes in developing countries.

Table 3 summarizes the economic consequences of the green revolution. Production increases are separated into increases from higher crop area planted and increases from higher yields. Yield increases are further separated into GRMV contributions and other input (fertilizer, labour) contributions. In the early green revolution period, production increased by 3.2 per cent a year. Yield increases account for 2.5 per cent a year. In the late green revolution period, production increased by 2.2 per cent per year. Yield increases accounted for 1.8 per cent per year. The sub-Saharan Africa region was an outlier in both periods, with low modern varieties

Agricultural Research, Table 2 Average annual varietal releases by crop and region, 1965–2000

Crop	1965–70	1971–75	1976–80	1981–85	1986–90	1991–95	1996–00
Wheat	40.8	54.2	58.0	75.6	81.2	79.3	80
Rice	19.2	35.2	43.8	50.8	57.8	54.8	58.5
Maize	13.4	16.6	21.6	43.4	52.7	108.3	71.3
Sorghum	6.9	7.2	9.6	10.6	12.2	17.6	14.3
Millets	0.8	0.4	1.8	5.0	4.8	6.0	9.7
Barley	0.0	0.0	0.0	2.8	8.2	5.6	7.3
Lentils	0.0	0.0	0.0	1.8	1.8	3.9	5.0
Beans	4.0	7.0	12.0	18.5	18.0	43.0	40.0
Cassava	0.0	1.0	2.0	15.8	9.8	13.6	14.0
Potatoes	2.0	10.4	13.0	15.9	18.9	19.6	20.0
All crops							
Latin America	37.8	55.9	65.9	92.5	116.2	177.3	139.2
Asia	27.2	59.6	66.8	86.3	76.7	81.2	79.9
Middle East–North Africa	4.4	8.0	10.2	12.2	28.4	30.5	82.2
Sub-Saharan Africa	17.7	18.0	23.0	43.2	46.2	50.1	55.2
All regions	87.1	132.0	161.8	240.2	265.8	351.7	320.5

Source: Evenson (2003a)

Agricultural Research, Table 3 Economic consequences of the green revolution (growth rates of food production, area, yield, and yield components, by region and period)

	Early green revolution 1961–80	Late green revolution 1981–2000
Latin America		
Production	3.083	1.631
Area	1.473	−0.512
Yield	1.587	2.154
MV contributions to yield	0.463	0.772
Other input/ha	1.124	1.382
Asia		
Production	3.649	2.107
Area	0.513	0.020
Yield	3.120	2.087
MV contributions to yield	0.682	0.968
Other input/ha	2.439	1.119
Middle East–North Africa		
Production	2.529	2.121
Area	0.953	0.607
Yield	1.561	1.505
MV contributions to yield	0.173	0.783
Other input/ha	1.389	0.722
Sub-Saharan Africa		
Production	1.697	3.189
Area	0.524	2.818
Yield	1.166	0.361
MV contributions to yield	0.097	0.471
Other input/ha	1.069	−0.110

(*continued*)

Agricultural Research, Table 3 (continued)

	Early green revolution 1961–80	Late green revolution 1981–2000
All developing countries		
Production	3.200	2.192
Area	0.683	0.386
Yield	2.502	1.805
MV contributions to yield	0.523	0.857
Other input/ha	1.979	0.948

Notes: Data on food crop production and area harvested are taken from FAOSTAT (2003) on total cereals, total roots and tubers, and total pulses. Asia: Developing Asia minus the countries of the Near East in Asia
Africa: Developing Africa minus the countries of the Near East in Africa and the countries of North-West Africa
Middle East–North Africa: Near East in Africa, Near East in Asia, and North-West Africa. Latin America: Latin America and the Caribbean
Crop production is aggregated for each region using area weights from 1981
Estimates of production increases due to MVs are from Evenson (2003b). Growth rates of other inputs are taken as a residual. Growth rates are compound and are computed by regressing time series data on a constant and trend variable. The totals for All developing countries are derived by weighting the regional figures by 1981 area shares
Source: Evenson and Gollin (2003)

(MV) contributions. The green revolution for sub-Saharan Africa was not accompanied by increased inputs, as it was in Asia and Latin America. (At least 12 countries – Afghanistan, Angola, Burundi, Central African Republic, Congo (Brazzaville), Gambia, Guinea Bissau, Mauritania, Mongolia, Niger, Somalia and Yemen – did not have a Green Revolution. Most are in sub-Saharan Africa.)

The Recombinant DNA (rDNA) Gene Revolution

In 1953 Watson and Crick published work (Watson 1968) that identified the 'double helix' structure of DNA and established DNA as the carrier of genetic information. In 1974 Cohen at Stanford and Boyer at the University of California at San Francisco achieved recombinant DNA 'transformation' or insertion of 'alien' DNA into organisms, and the field of genetic engineering was born (Cohen 1997).

Within a few years many 'crop biotech' companies were established. Large agricultural chemical companies were early entries into the field. Today seven life science firms (Monsanto, DuPont, and Dow in the US, Syngenta, BASF, and Bayer in Europe, and Savia in Mexico) dominate the genetically modified (GM) crop products industry. The first GM products introduced in the late 1980s were commercial failures. But bovine somatotrophin hormone (BsT), a product to stimulate milk production, was successfully introduced in 1993.

In 1995 several companies introduced GM crop products for canola (rapeseed), soybeans, maize and cotton. These products fall into two classes: herbicide tolerance and insect resistance *(Bacillus thuriengensis*, B_T). Herbicide tolerance (soybeans, canola and maize) enables weed control with traditional herbicides. This trait has been highly valued by farmers and rapidly adopted. Most of the world's canola and soybeans now have this trait, as does considerable acreage of maize. Insect resistance is achieved by engineering maize and cotton plants to produce B_T toxins that limit insect damage to the plant. This has a particularly important effect on cotton, where insects cannot readily be controlled by insecticides.

GM crop products enable farmers to reduce production costs. Cost reductions depend on mechanization status and insect pest status. Estimates of cost reduction vary by country, with Western European countries having negligible cost reduction potential (less than one per cent, because they produce little cotton, canola or soybeans). The US has significant cost reduction

Agricultural Research, Table 4 Returns to agricultural research studies

	No. of IRRs	Distribution of internal rates of return (% Median IRR distribution)						Median IRR
		0–20	21– 40	41–60	61–80	81–100	100 +	
Project evaluation methods	121	.25	.31	.14	.18	.06	.07	40
Statistical methods	254	.14	.20	.21	.12	.10	.20	50
Aggregate programmes	126	.16	.27	.29	.10	.09	.09	45
Pre-invention science	12	0.00	.17	.38	.17	.17	.17	60
Private sector R&D	11	.18	.09	.45	.09	.18	0.00	50
By region								
OECD	146	.15	.35	.21	.10	.07	.11	40
Asia	120	.08	.18	.21	.15	.11	.26	67
Latin America	80	.15	.29	.29	.15	.07	.06	47
Africa	44	.27	.27	.18	.11	.11	.05	37

Source: Evenson (2001)

potential, as do many developing countries. It should be noted, however, that cost reduction gains are 'static' in nature (that is, they do not cumulate over time). Dynamic gains can be produced only by the development of generations of modern varieties, as reflected in Table 2 for GRMVs. The gene revolution is not a substitute for the green revolution.

The gene revolution has become strongly politicized in recent years. A clear division has emerged between the original European Union countries and North American countries. The European Union position is that the 'precautionary principle' should apply, while the North American position is that, in the absence of scientific evidence to the contrary, farmers should be allowed to adopt GM crops (see FAO 2004).

Returns to Agricultural Research

Griliches (1958) was the first economist to measure 'returns to research' by computing returns to hybrid corn research. To do this, he created a cost stream and a benefit stream, and applied present value methods to them. (At a five per cent discount rate the present value of benefits was roughly seven times the present value of costs. Some interpreted this as a 700 per cent rate of return. Of course, it was in fact a benefit–cost ratio.) Griliches computed an internal rate of return to hybrid corn research of 43 per cent.

Agricultural Research, Table 5 Green revolution returns to research

Countries	IARCs	NARS
Latin America	39	31
Asia	115	33
West Asia–North Africa	165	22
Sub-Saharan Africa	68	9

Source: Evenson (2003b)

Evenson (2001) reviewed more than 300 studies of returns to research in the decades after the Griliches studies. Table 4 reports a summary of internal rates of return reported in these studies. The project evaluation studies utilized methods similar to those used by Griliches. The statistical studies generally regressed measures of total factor productivity on research stock variables. Some studies were focused on specific commodities, others on aggregate research programmes. Several studies made a distinction between pre-invention science and applied science, and several studies were undertaken of the private sector contribution to agriculture.

The studies are characterized by great diversity in internal rates of return (IRRs), ranging from IRRs of zero to very high levels. Median IRRs are high for all categories. This diversity is consistent with the fact that research is a highly uncertain activity.

Finally, Table 5 utilizes data from the green revolution where GRMV adoption rates were

available. The method applied was similar to that which Griliches originally used. These data confirm the estimates in Table 4. Very high returns to IARC research are shown. Returns to NARS programmes are lower, especially in sub-Saharan Africa where many countries did not achieve a green revolution.

See Also

- ▶ Agriculture and Economic Development
- ▶ Population and Agricultural Growth
- ▶ Technology

Bibliography

Cohen, J. 1997. The genomics gamble. *Science* 275: 767–772.

Ehrlich, P. 1968. *The population bomb*. New York: Sierra Club/Ballantine Books.

Evenson, R. 2001. Economic impacts of agricultural research and extension. In *Handbook of agricultural economics*, ed. B. Gardner and G. Rausser, Vol. 1. Amsterdam: Elsevier Science B.V.

Evenson, R. 2003a. Modern variety production: a synthesis. In *Crop variety improvement and its effect on productivity: The impact of international research*, ed. R. Evenson and D. Gollin. Wallingford: CABI Publishing.

Evenson, R. 2003b. Production impacts of crop genetic improvement. In *Crop variety improvement and its effect on productivity: The impact of international research*, ed. R. Evenson and D. Gollin. Wallingford: CABI Publishing.

Evenson, R., and D. Gollin. 2003. Assessing the impact of the green revolution, 1960 to 2000. *Science* 300: 758–762.

FAO (Food and Agriculture Organization of the United Nations). 1998. *The state of the world's plant genetic resources for food and agriculture*. Rome: FAO.

FAO. 2004. *The state of food and agriculture, 2003–2004*. Rome: FAO.

FAOSTAT. 2003. Agricultural data. FAO statistical databases. Online. Available at http://apps.fao.org/page/collections?subset=agriculture. Accessed 15 Sep 2005.

Griliches, Z. 1957. Hybrid corn: An exploration in the economics of technical change. *Econometrica* 25: 501–522.

Griliches, Z. 1958. Research costs and social returns: Hybrid corn and related innovations. *Journal of Political Economy* 66: 419–431.

Watson, J. 1968. *The double helix: A personal account of the discovery and structure of DNA*. New York: Athenium.

Agricultural Supply

Jere R. Behrman

One of the earliest-investigated and most fruitful areas for econometric studies has been the estimation of agricultural supply functions. Studies date back at least to the work of Smith (1928) and Bean (1929) on US agriculture in the 1920s. Early studies adopted a fairly static view. Nerlove's (1958) work on *The Dynamics of Supply*, with adaptive price expectations and adjustment processes for United States agriculture, spawned renewed interest in specification and estimation issues on this topic. The roughly concurrent controversies about market responsiveness in developing-country agriculture (e.g., Schultz 1964) led to a shift in emphasis towards this concern. In the last quarter century a veritable flood of such studies has appeared. More recently supply studies have incorporated more systematic emphasis on systemic characteristics, risk aversion, the household/farm model framework and alternative price expectations. This article reviews these basic developments in empirical studies of agricultural supply by starting with the most common framework for such analyses and then considering what questions arise when some of the traditional assumptions are weakened.

Perfectly-Competitive, Equilibrium Supplies with No Risk

Most empirical studies of agricultural supply have assumed perfect competition, equilibrium, no risk and separability between the farm production decisions and the farm household consumption decisions. Under such conditions the supply function for a specific product of an individual producer is the marginal cost function for product prices sufficiently high so that variable costs are covered (and zero otherwise). The market supply function is the sum of all individual supply functions. As such, the supply function depends on all

relevant expected product and input prices, all fixed factors, and technology. There are two important elements of dynamics in the supply process: the adjustment of short-run fixed factors over longer time periods and the creation of expectations for product harvest prices at the time that inputs, especially land, are committed.

Early studies basically posited a supply function as dependent on relevant expected prices (P^*):

$$S = f(P^*) \tag{1}$$

where S is a vector of supplies of different agricultural products and P^* is a vector of expected (or, at least after Nerlove's contributions, expected normal) prices.

Expected prices most commonly were represented by one-period lagged prices for products for which actual harvest prices would not be known until harvest time and by actual prices at the time of the input decision for inputs. Frequently, instead of using supplies as the dependent variable, areas devoted to individual crops were used since land is a critical input for which allocation among crops basically is under the control of the farmer.

The seminal contributions of Nerlove (1956, 1958) were: first to generalize the formation of price expectations with adaptive expectations and second, to incorporate a distributed-lag adjustment process to reflect that adjustment is not costless. His basic model for an annual crop, thus, is:

$$A_t^* = a_0 + a_1 P_{\mathrm{t}}^* + a_2 Z_t + u_t \tag{2}$$

$$A_t = A_{\mathrm{t}-1} + \gamma\left(A_t^* - A_{t-1}\right) \tag{3}$$

$$P_{\mathrm{t}}^* = P_{\mathrm{t}-1}^* + \beta\left(P_{t-1} - P_{t-1}^*\right) \tag{4}$$

where

A_{t} is actual area under cultivation in period t,
A_{t}^* is 'desired' or equilibrium area under cultivation in t,
P_{t} is actual product price in t,
P_t^* is 'expected normal' price in t,
Z_{t} is other observed, exogenous factors, and
u_{t} is a disturbance term.

Relation (Eq. 2) states that desired area is a function of expected normal prices, other exogenous factors, and a disturbance term; but neither the desired area nor the expected price typically is observed. Relation (Eq. 3) states that there is a distributed-lag adjustment in actual area towards desired area, with γ the 'coefficient of adjustment'. Relation (Eq. 4) states that the expected normal price in period t is the expected normal price in the previous period plus an adjustment for the discrepancy in the previous period between the expected normal and the actual price. Substitution of Eqs. 3 and 4 into Eq. 2 gives an expression in terms of observable variables:

$$\begin{aligned} A_t = {} & [(1-\beta) + (1-\gamma)]A_{\mathrm{t}-1} \\ & - (1-\beta)(1-\gamma)A_{\mathrm{t}-2} + a_1\gamma\beta P_{\mathrm{t}-1} \\ & + a_2\gamma Z_t - a_2\gamma(1-\beta)Z_{\mathrm{t}-1} \\ & + a_0 + \gamma u_t - \gamma(1-\beta)u_{\mathrm{t}-1}. \end{aligned} \tag{5}$$

Literally hundreds of estimates of some variant of this supply relation have been made. Initially they focused on developed-country agriculture. Due to the debate over the market-responsiveness of traditional developing agriculture, emphasis then shifted relatively to developing-country agriculture in studies by Krishna (1963, 1965), Behrman (1966, 1968b) and a host of others. Askari and Cummings (1976, 1977) noted over 600 such estimates by the mid 1970s. These studies vary substantially regarding the identification of relevant observed exogenous variables, the treatment of serial correlation in the disturbance term, what prices are included, and the role of yields. Despite such variations they point to a pattern of significant and often substantial price responses in agricultural supplies, with some indication of greater responses for higher-income and more-literate farmers, larger farms, own-operated farms, farms with access to irrigation, and crops with lower yield variability. This price responsiveness in developing-country agriculture suggests that measures to suppress particular agricultural prices significantly discourage domestic agricultural supplies of those products. An aggregate supply response study for a cross-section of developing countries by Peterson (1979) also

reports substantial discouragement of aggregate agricultural production in these countries by price policies, though Binswanger et al. (1985) suggest that this result is an artifact of the price data used. The price responsiveness in developed countries suggests that price and income-support programmes induce expanded agricultural supplies for the products affected.

One subset of these studies merits particular mention: those that relate to perennials and livestock. Most of these studies have adapted the above model to incorporate long lags due to the long gestation between investment and production, with adjustment lags posited in respect to desired capital stock or desired investment. In a few cases (e.g. French and Matthews 1971) there are extensive explicit empirical representations of the various critical variables in perennial and livestock production: production, investment, non-bearing new capital, and old capital removal. In many cases, however, the lack of basic data on capital stocks has left bearing stocks to be inferred from outputs or controlled for by differencing outputs for products such as cocoa in which there is a long period once trees have matured during which yields are approximately constant (e.g., Bateman 1965, 1968; Behrman 1968a, 1969). Despite the greater complexities of perennials and livestock production and the greater longevity of the related capital stock, there has been little effort to go beyond an essentially static formulation for the demand for the relevant capital stock. Conceptually this problem can be formulated as a dynamic programming model. But, as Nerlove (1979) notes, severe difficulties exist in the empirical implementation of such a strategy because of data inadequacies and because of the uncertainty of future technological developments.

Two additional modifications subsequent to the widespread use of the supply relation in Eq. 5 are worth noting. First, the more recent formulation of the supply relation is to start with a profit function and to note that the partial derivative with respect to a particular product price gives the supply function for that product and the partial derivative with respect to the price for a production input gives the input demand function. This approach has the advantage of focusing on the interrelations in a system of supply and demand relations within a multi-product and multi-input context and in indicating the nature of cross-equation systemic restrictions. The natural distinction between outputs and inputs within this context also sharpens the question about the frequent usage of area as the dependent variable in supply-response relations; such estimates presumably are characterized better as approximations to input demand relations though, as such, they provide information on the underlying parameters of the system that pertain to supply responses. Examples of profit-function-based agricultural supply studies include Lau and Yotopoulos (1971) and Bapna et al. (1984). In most applications of this approach, the concerns of Nerlove (1956, 1958) about the empirical dynamics of price expectation formulation and of adjustment have been ignored. Instead, assumptions of immediate adjustment and static product price expectations (i.e., the previous period's prices) prevail. There would seem to be further gains in understanding from the incorporation of such dynamic concerns into these system estimates, though full incorporation of such dynamics leads to dynamic programming models with the problems mentioned above.

Second, representations of price expectations have changed. There has been growing emphasis throughout economics on 'rational expectations' (Muth 1961), which are the minimum mean square forecasts based on the information available at the time of the forecast, including that about the structure of the system. There have been some efforts to incorporate rational price expectations into agricultural supply studies, though with reduced-form relations with the same variables as in relation (Eq. 4) (e.g., Eckstein 1984). The question remains open whether such rational expectations are preferable representations of expectations actually held by farmers and, if they are, about what information is available for farmers to utilize in their expectation formulation. Nerlove (1979) observes, for example, that the rejection of the proposition that farmers respond to the expectation of some average of prices in all future periods with adaptive expectations also implies the rejection of the notion that farmers are adjusting to a well-defined, longer-run equilibrium because such an equilibrium is well-defined only for stationary price expectations.

Non-separability Between Farm Production and Consumption

For the currently developed countries, the separability assumption probably is plausible. Most production is sold on markets, most consumption goods are purchased on markets, there probably are not consumption–labour productivity links and markets are relatively complete (though risk may be a problem, see below).

For the developing countries, in contrast, the separability assumption may be misleading for millions of farm households. One respect in which these assumptions may be misleading is with regard to responses in total production versus the marketed surplus. Many of these households consume large shares of the basic staples that they produce. As a result, price responses in the marketed surplus may differ substantially from those in total production, depending on what is the household own-consumption response. Krishna (1965) and Behrman (1966) presented early models of the price elasticity of the marketed surplus that incorporate the income elasticity of consumption within the household. Such models demonstrate that the price elasticity of the marketed surplus may differ greatly from those of total supply if own-consumption accounts for a substantial share of production and if either own-consumption price or income elasticities is large.

Two other reasons for which integrated household-farm models have been emphasized are incomplete markets (e.g., Lau et al. 1978; Barnum and Squire 1979a, b) and productivity–consumption links (e.g., Leibenstein 1957; Bliss and Stern 1978a, b; Stiglitz 1976; Pitt and Rosenzweig 1986; Behrman and Deolalikar 1987; Strauss 1986). Both these phenomena are thought to be common in developing countries. If they are important, agricultural supply should be explored within the context of the household-farm model. This means that prices for consumption goods and services (e.g., for schools, clothing and health-care) and fixed household assets should enter into the determination of agricultural supplies, in addition to prices of agricultural products and inputs and fixed agricultural factors. For the examination of the determinants of some perennials and livestock, for instance, even the prices that determine births, infant and child mortality and migration in principle should be included since such prices simultaneously determine the long-run agricultural capital stocks (with the obvious link through the long-run availability of household labour). While there are a few studies of agricultural supply that have emphasized the conceptual importance of the farm-household model (e.g., Lau et al. 1978; Barnum and Squire 1979a, b), in empirical applications the full ramifications of such demand-production simultaneity are yet to be explored.

Risk and Risk Aversion

Farmers are subject to production risk and, for farmers who partake in markets, price and input-availability risks. Once it was established that developing-country farmers seem responsive to expected prices, considerable emphasis shifted to the role of risk and risk aversion in determining agricultural supplies. Many studies have attempted to test for the supply response to risk by including ad hoc empirical measures of risk (e.g., variances in prices or in yields as in Behrman 1968b) and report some evidence of negative responses to risk. Several experiments have been undertaken to attempt to identify the nature and the magnitude of risk aversion among farms. The most satisfactory of these to date is by Binswanger (1980, 1981) in which the payoffs were real and substantial. He concludes that his results are consistent with expected utility maximizing behaviour (and not with security-based forms of behaviour in which farmers are concerned primarily with avoiding disaster) and that most individuals are risk averse, but not very risk averse.

How should supply responses be modelled given the possibility of risk and risk aversion? Newbery and Stiglitz (1981) provide a recent theoretical synthesis of the implications of risk and risk aversion for modelling supply in their discussion of the theory of commodity price stabilization. They demonstrate that risk may have an impact even on a risk-neutral farmer; such a farmer does not just maximize the product of

expected prices and expected quantities minus costs if prices and quantities are correlated (as would be the case for a perfectly competitive farmer in an area which accounts for a large share of the market, as with West African cocoa production), but also must incorporate the price-quantity covariance in order to maximize expected profits. They also argue that rigorous specification of supply behaviour under risk aversion is difficult and should proceed from first principles of constrained utility maximization (which is likely to require a farm-household framework as well). Binswanger (1982) further elaborates on the difficulties of econometric estimation under risk preferences. I am unaware of empirical studies to date that are consistent with such a framework.

Disequilibrium

The standard framework for empirical agricultural supply analysis assumes equilibrium, or adjustment towards equilibrium in which observed prices convey all the available information. As Schultz (1975) and Nerlove (1979) emphasize, however, for some important questions such as the nature of the historical transformation of developed-country agriculture and the current transformations of developing-country agriculture, disequilibria are likely to be common and visible prices are not likely to convey all of the relevant information available to farmers. For studies of agricultural supply responses in such contexts, a broader perspective is desirable to represent the impacts of differential capabilities of economic entities to deal with disequilibria, public investments, development of markets, technological and demographic changes, and governmental roles. Embedding supply studies within this larger context in order to attain further understanding remains a major challenge.

See Also

- ▶ Adaptive Expectations
- ▶ Agricultural Economics
- ▶ Cobweb Theorem
- ▶ Production and Cost Functions

References

Askari, H., and J.T. Cummings. 1976. *Agricultural supply response: A survey of the econometric evidence*. New York: Praeger Publishers.

Askari, H., and J.T. Cummings. 1977. Estimating agricultural supply response with the Nerlove model: A survey. *International Economic Review* 18: 257–292.

Bapna, S.L., H. Binswanger, and J.B. Quizon. 1984. Systems of output supply and factor demand equations for semi-arid tropical India. *Indian Journal of Agricultural Economics* 39(2): 179–202.

Barnum, H.N., and L. Squire. 1979a. An econometric application of the theory of farm-household. *Journal of Development Economics* 6(1): 79–102.

Barnum, H.N., and L. Squire. 1979b. *A model of an agriculture household: Theory and evidence*. Baltimore: Johns Hopkins for the World Bank. chs 3–6.

Bateman, M. 1965. Aggregate and regional supply functions for Ghanaian cocoa, 1946–62. *Journal of Farm Economics* 47: 384–401.

Bateman, M. 1968. *Cocoa in the Ghanaian economy: An econometric model*. Amsterdam: North-Holland.

Bean, L.H. 1929. The farmers' response to price. *Journal of Farm Economics* 11: 368–385.

Behrman, J.R. 1966. Price elasticity of the marketed surplus of a subsistence crop. *Journal of Farm Economics* 48: 875–893.

Behrman, J.R. 1968a. Monopolistic cocoa pricing. *American Journal of Agricultural Economics* 50: 702–719.

Behrman, J.R. 1968b. *Supply response in underdeveloped agriculture: A case study of four major annual crops in Thailand 1937–1963*. Amsterdam: North-Holland.

Behrman, J.R. 1969. Econometric model simulations of the world rubber market 1950–1980. In *Essays in industrial econometrics*, vol. 3, ed. L.R. Klein. Philadelphia: Economic Research Unit, Wharton School of Finance and Commerce, University of Pennsylvania.

Behrman, J.R., and A.B.. Deolalikar. 1987. Health and nutrition. In *Handbook on development economics*, ed. H.B. Chenery and T.N. Srinivasan. Amsterdam: North-Holland.

Binswanger, H.P. 1980. Attitudes towards risk: Experimental measurement evidence in rural India. *American Journal of Agricultural Economics* 62(3): 395–407.

Binswanger, H.P. 1981. Attitudes towards risk: Theoretical implications of an experiment in rural India. *Economic Journal* 91: 867–890.

Binswanger, H.P. 1982. Empirical estimation and use of risk preferences: Discussion. *American Journal of Agricultural Economics* 64(2): 391–393.

Binswanger, H., Y. Mundlak, M.C. Yang, and A. Bowers. 1985. *Estimation of aggregate supply response*. Washington, DC: World Bank. Report No. ARU 48.

Bliss, C., and N. Stern. 1978a. Productivity, wages and nutrition, part I: The theory. *Journal of Development Economics* 5(4): 331–362.

Bliss, C., and N. Stern. 1978b. Productivity, wages and nutrition, part II: Some observations. *Journal of Development Economics* 5(4): 363–398.

Eckstein, Z. 1984. A rational expectations model of agricultural supply. *Journal of Political Economy* 92(1): 1–19.
French, B.C., and J.L. Matthews. 1971. A supply response model for perennial crops. *American Journal of Agricultural Economics* 53: 478–490.
Jarvis, L. 1974. Cattle as capital goods and ranchers as portfolio managers: An application to the Argentine cattle sector. *Journal of Political Economy* 82: 489–520.
Krishna, R. 1963. Farm supply response in India–Pakistan: A case study of the Punjab region. *Economic Journal* 73: 477–487.
Krishna, R. 1965. The marketable surplus function for a subsistence crop: An analysis with Indian data. *Economic Weekly* 17.
Lau, L.J., and P.A. Yotopoulos. 1971. A test for relative efficiency and application to Indian agriculture. *American Economic Review* 61: 94–109.
Lau, L.J., W.-L. Lin, and P. Yotopoulos. 1978. The linear logarithmic expenditure system: An application to consumption-leisure choice. *Econometrica* 46(4): 843–868.
Leibenstein, H. 1957. *Economic backwardness and economic growth*. New York: Wiley.
Muth, J.F. 1961. Rational expectations and the theory of price movements. *Econometrica* 29: 315–335.
Nerlove, M. 1956. Estimates of supply of selected agricultural commodities. *Journal of Farm Economics* 38: 496–509.
Nerlove, M. 1958. *The dynamics of supply: Estimation of Farmers' response to price*. Baltimore: Johns Hopkins University Press.
Nerlove, M. 1967. Distributed lags and unobserved components in economic time series. In *Ten economic essays in the tradition of Irving fisher*, ed. W. Fellner et al., 126–169. New York: Wiley.
Nerlove, M. 1979. The dynamics of supply: Retrospect and prospect. *American Journal of Agricultural Economics* 61(5): 867–888.
Newbery, D.M.G., and J.E. Stiglitz. 1981. *The theory of commodity price stabilization*. Oxford: Clarendon Press.
Nowshirvani, V. 1971. Land allocation under uncertainty in subsistence agriculture. *Oxford Economic Papers* 23: 445–455.
Peterson, W.L. 1979. International farm prices and the social cost of cheap food policies. *American Journal of Agricultural Economics* 61: 12–21.
Pitt, M.M., and M. Rosenzweig. 1986. Agricultural prices, food consumption and the health and productivity of farmers. In *Agricultural household models: Extensions, applications and policy*, ed. I.J. Singh, L. Squire, and J. Strauss. Washington, DC: World Bank.
Schultz, T.W. 1964. *Transforming traditional agriculture*. New Haven: Yale University Press.
Schultz, T.W. 1975. The value of the ability to deal with disequilibrium. *Journal of Economic Literature* 13: 827–846.
Smith, B.B. 1928. *Factors affecting the price of cotton*. US Department of Agriculture Technical Bulletin No. 50, Washington, DC.
Stiglitz, J. 1976. The efficiency wage hypothesis, surplus labour, and the distribution of income in LDC's. *Oxford Economic Papers, New Series* 28: 185–207.
Strauss, J. 1986. Does better nutrition raise farm productivity? *Journal of Political Economy* 94(2): 297–320.

Agriculture and Economic Development

John W. Mellor

Abstract

Agriculture plays a vital role in economic development by facilitating the transition from a low-income subsistence to a high-income commercial economy. Agriculture promotes economic transformations by supplying food, foreign exchange, labour, and effective demand to the non-farm sectors, and is the dominant force in poverty reduction. A land constraint makes agricultural growth unusually dependent on technological change, while geographically dispersed production units favour a family-size labour force. These in turn lead to a special role for government in achieving rapid agricultural growth.

Keywords

Agriculture and economic development; Economic development; Economic growth; Employment growth; Family; Food and Agriculture Organizational (FAO); Foreign aid; Government role; Green Revolution; International Food Policy Research Institute (IFPRI); Non-tradable commodities; Poverty reduction; Productivity; Protection; Research expenditure; Rural non-farm sector; Rural-urban income disparities; Technological change

JEL Classifications

O13

Economic development is characterized by three transformations: from domination by agriculture to domination by manufacturing and then services; from domination by non-tradable goods and services to a much larger weight of tradable goods and services; and, from a high proportion of poor people living at the edge of basic subsistence to one with few or no such people. If those transformations are to proceed rapidly and efficiently, agriculture must play a vital role. In the course of playing that role, the relative size of agriculture declines drastically while its absolute size increases.

Agriculture has several characteristics that define not only its ability to influence the various transformations but also the means by which it grows and facilitates those transformations. The most important of these are threefold: first, dependence on land and a land constraint that yields rapidly diminishing returns to increased inputs, making agriculture unusually dependent on technological change for its growth; second, geographically dispersed production units that favour a family-size labour force, with the amount of land and capital per family increasing immensely with rising incomes; and third, derived from the first two, a special role for government in meeting the conditions of rapid agricultural growth. Reinforcing the need for good governance is the increasing need for government-provided institutions for ensuring a healthy, educated labour force as agriculture modernizes.

The Size of Agriculture

Initially humankind produced the basic means of substance at such low levels of productivity that there was time for little else. Agriculture dominated those subsistence activities. From that initial base, progress could be made only by increased productivity in agriculture, thereby releasing resources for other needs and eventually for luxuries. Even in lower middle-income countries agriculture remains sufficiently large that it continues to play a critical role in transforming the economy. Agriculture's role in employment growth, raising real wage rates and hence reducing poverty is even greater than its role in GDP growth. It continues to be dominant in employment growth at least through upper middle-income status.

Share of GDP

In low-income countries, such as those in most of contemporary Africa, significant parts of Latin America, and, until recently, most of Asia, agriculture accounts for in the order of half of GDP. By the time middle-income status is reached, as it has been in most of contemporary Asia, Latin America and the Middle East, agriculture's relative importance has declined to between 15 and 25 per cent of GDP. With high-income status it declines to under five per cent.

However, as the economy is transformed agriculture can still grow rapidly in absolute size. Indeed, the faster agriculture grows in absolute terms the faster its relative importance declines. This is because high-income elasticity of demand by farmers for non-farm goods and services causes those sectors to grow faster than agriculture and all the more so at high rates of agricultural growth (see Mellor 1995).

The decline in the relative size of agriculture is further hastened by the appearance of scale economies in many of the production and marketing services for modern agriculture. As development proceeds, many tasks performed on farms in the early stages of development are more economically produced by large-scale firms. Initially farmers produced their own plant nutrients from composting and manure, but it became much cheaper to buy inorganic fertilizers from immense petrochemical plants. Power initially is derived from humans and animals raised on the farm but eventually from tractors and other machines produced off the farm. The examples are endless.

Share of Employment and Employment Growth

Statistics for agriculture's share of employment in low- and middle-income countries are always far larger than those for its share of GDP. That is substantially because of misclassification. Persons with very small holdings that are insufficient in size to provide even half of family employment or

income are normally classified as farmers, but of course they are more properly classified as rural non-farm population given the way they make their living. Thus, typically even in low-income countries the rural population is divided about equally between those who make their living primarily in farming and those in other rural occupations. Seen this way, farmers represent a similar proportion of employment and GDP. This is not surprising since farm income derives substantially from land ownership, not just from labour, just as in the urban sector income derives substantially from return on capital as well as from labour.

Thus, in a low-income country 80–90 per cent of the population may be rural, half with farming as their principal occupation. By the time middle-income status arrives the share of population that is rural has declined to around 40 per cent and the share principally occupied in farming to less than 20 per cent. In high-income countries the farm population is less than five per cent, two-thirds of those producing the bulk of the farm output.

Agriculture and Economic Growth

Because of its initially dominant size, agriculture makes several large initial contributions to overall growth (Mellor 1976.) Growth in agricultural productivity releases labour for the fast-growth non-farm sectors. Agriculture earns foreign exchange that is utilized to import capital goods for the non-farm sector. It provides low-cost food to keep labour costs down as employment in the non-farm sector grows rapidly. Rapid growth in non-farm employment faces rapidly rising, competitiveness-destroying increases in real wage rates if agricultural production does not grow rapidly. Even in an open economy with rapid growth in urban incomes, increased food imports would be so great with a failing agriculture that the real exchange rate would change sharply and push up the cost of food and therefore of labour. Finally, fast-growth agriculture plays the dominant role in employment growth and poverty reduction. In the context of modern open economies and free capital flows, the latter contribution remains the most important for agriculture.

Agriculture and Poverty Reduction

Statistical data from diverse cross-sectional analyses show that in low- and middle-income countries it is agricultural growth that drives poverty reduction (Ravallion and Datt 1996.) Further, there is a significant lag in that poverty-reducing impact. The lack of immediate impact led to an incorrect view that agricultural growth does not reduce poverty. It is now known that the lag is due to the large indirect impact of agricultural growth on poverty reduction. There is, however, a major exception to this relation. When land ownership is highly skewed, as for example in much of Latin America, agricultural growth does not significantly reduce poverty. That is because very rich people with large landholdings spend additional income not on employment-intensive rural non-farm goods and services but on capital and import-intensive urban goods and services.

When agricultural incomes are broadly distributed, agricultural growth reduces urban poverty more than does urban growth. This is because urban poverty is a product of rural-urban wage disparities. If rural incomes are stagnant, the rural-urban disparity increases and poor rural people migrate to the cities. If the disparity is large, rural people will be willing to wait a long time in the urban area for a job, living in poverty in slums. The return to waiting is made up once they get the good urban job. Thus, the greater the income disparity, the longer the queue and hence the greater the number in urban poverty. Measures that make waiting cheaper, such as subsidized housing or even normal urban amenities such as potable water, simply increase the rural-urban disparity and hence the queue. Thus, the way to reduce urban poverty is to raise rural incomes and amenities as rapidly as those in urban areas.

There are three means by which agricultural growth contributes to reduced poverty: lower food prices; increased agricultural employment; and farm income-driven rural non-farm employment.

Food Prices

Poor people in low-income countries spend in the order of 80 per cent of their income on food. It follows that the real price of food is a primary

determinant of the real income of the poor. In a neoclassical economy, increased domestic food production does not reduce the price of food because the international price rules. However, high transfer costs in low-income countries somewhat insulate domestic food prices from international prices. This may be reinforced by trade restrictions. In that case, increasing food production faster than domestic demand will reduce domestic food prices and greatly benefit poor people. The high-yielding rice varieties that brought the Green Revolution to Asia were of low quality, depressing the price of rice consumed by the poor. Market forces may depress the nominal wage as food prices decline, but those same market forces will then increase employment. Hence, the poor tend to benefit from rapid growth in agriculture either through lower food prices or through increased employment (Mellor 1976).

Of course, these same processes work in reverse. If agricultural production grows more slowly than domestic demand, food prices tend to rise, reducing the real incomes of the poor. Unfavorable weather reduces agricultural production; prices rise and the poor suffer. Wage rates rarely adjust in the short run, although they do in the long run, in which case higher wage rates reduce employment. In either case the poor lose.

Of course, increasing the supply of food faster than demand is difficult in low-income countries in which population growth is rapid and in which incomes may also be rising. The income elasticity of demand for food is much less inelastic in low-income countries than in high-income countries, and hence income growth has a major effect on the demand for food. For example, the Food and Agriculture Organizational of the United Nations (FAO) and the International Food Policy Research Institute (IFPRI) both show for Africa continued shortfall in supply into the indefinite future. The African poor will continue to suffer from such trends (Eicher and Staatz 1998).

Increased Farm Employment

Because agriculture is initially so large, rapid growth does add directly and substantially to employment. However, direct employment growth is small compared with the growth in output. This is because productivity-increasing technological change is the primary source of high growth rates in agriculture. Even though the technology is generally designed to be land-saving, it also increases labour productivity. Thus, for each ten per cent increase in agricultural output employment increases by between less than three per cent and at most six per cent. Thus, the big impact of agricultural growth on employment comes indirectly through the rural nonfarm sector.

Increased Rural Non-farm Employment: Driven by Rising Farm Incomes

In an open economy, agricultural output that grows faster than demand does not depress prices significantly because of access to international markets. A small decrease in prices brings increased exports. A high growth rate in output without depression of prices raises farm income and reduces poverty in a quite different manner from that of reduced prices.

Farmers spend a large and increasing proportion of increments to their income on the goods and services produced by local, rural non-farm workers. Numerous studies show that the bulk of the poor are rural non-farm workers. They largely produce non-tradable goods and services. Because of low quality and high transaction costs they cannot export as an alternative to meeting local demand.

When farmers prosper they enlarge their homes and buy local furniture, local tailoring, and a vast panoply of services. That increases employment and eventually real wages in the rural non-farm sector. This is the source of poverty reduction in a low- or medium-income open economy.

Because of the strong multiplier on those expenditures, there is a significant lag in the full effect of agricultural growth on poverty reduction as successive rounds of expenditure occur. Similarly, rich, and especially absentee, landowners spend incremental income largely on capital and import-intensive commodities and services and so have little effect on poverty reduction. These two relations are consistent with the data cited earlier.

In very poor agricultures that are growing little or not at all, those in the rural non-farm sector are

exceedingly poor because of lack of local demand. In that situation outmigration of the principal male worker and sending back of remittances are important factors holding poverty in check. This is, of course, a socially disruptive means of holding off poverty. Thus it is not surprising that when farm incomes rise rapidly migration beyond commuting range is sharply reduced.

Rural-Urban Income Disparities

It is not uncommon for the urban sector to grow rapidly in low-income countries, even while agriculture stagnates. Foreign aid may be spent largely in the cities, as in Africa, or macro policy stimulates manufacturing growth while the complex processes of agricultural growth are neglected. In that case, urban and rural poverty both surge. At that stage of development it is critical that agricultural production grows rapidly in order to prevent rapid widening of rural urban disparities.

As countries move to middle-income status, the problem of rural-urban disparities changes. The rate of growth of urban incomes accelerates – to around six per cent per year. The capacity to absorb migration also increases as the urban proportion of the population increases. Concurrently, the potential for accelerating the agricultural growth rate improves. The demand for high-value agricultural commodities, such as livestock products and fruits and vegetables, grows at a rate of between six and eight per cent, much of which can be efficiently met from domestic production. Thus, in middle-income countries the agricultural growth rate may pick up to between four and six per cent. That would allow rural incomes to roughly keep pace with urban incomes. While not uncommon amongst middle-income countries, such growth rates are by no means universal and require carefully selected government actions.

Characteristics of Agriculture That Determine the Means of Growth

Agriculture has very different characteristics from urban industry and therefore different requirements for growth (Eicher and Staatz 1998). If those divergent characteristics are not recognized then not only does agriculture grow slowly but poverty reduction halts and income disparities between rural and urban areas widen. A family-size labour force, the importance of technological change and rural infrastructure, and the consequent importance of government are the dominant characteristics that distinguish the process of agricultural growth from that of other sectors.

The most obvious characteristic of agriculture is that each farm is spread over a wide area. This disperses the workforce and, combined with the complex biological nature of the production process, puts a premium on family-size operating units (commonly including one hired worker) with minimal supervision costs. Size of farm measured by land area or capital investment varies immensely among countries; but the labour force per farm is a virtual constant.

Particularly in low- and middle-income countries in which both land and capital holdings are small as well, the small-scale unit requires support from activities with scale economies. Most of these activities are most efficiently pursued by the private sector. But some are public goods and require public sector activity.

The balance between public and private sectors gradually shifts towards the private sector as development occurs and the private sector cultivates a broadened set of skills. Particularly in low-income countries, such a substantial burden falls on the public sector, in research, extension, enforcement of grades and standards (especially for export), and some aspects of finance and of market information systems, that the government must set difficult priorities. In that context it must continually press to turn activities over to, and encourage, the private sector as that sector's capacity increases.

The key role of government in agricultural growth, in turn, makes the role of the agriculture ministry important as it diagnoses needs and facilitates and complements the private sector. Particularly in early stages of accelerated agricultural growth, the agriculture ministry must have an explicit strategy with clear priorities and

sequences in which to take up key activities. When the fashion in development swings towards minimizing the role of government, agriculture is more likely to suffer than other sectors.

Key Forces in Agricultural Growth

Much of what is required for rapid agricultural growth is most appropriately and efficiently undertaken in the private sector, but even the minimum set of required public sector activities is long and complex. Government can do only a few major things at a time. Thus, one of the most important elements of a high growth rate is an at least implicit strategy within which a small number of limiting priorities will be set with an efficient sequence of activities guiding the moving on to new priorities as earlier ones are fulfilled and institutionalized.

The immediate priorities differ from country to country depending on the physical circumstances and the history of interventions. Hence, setting priorities and sequences and even the broad strategy are highly country-specific. A few generalizations are possible. Physical infrastructure and technology institutions are critical in all growth plans, and government is essential to the provision of both. They are also both never-ending tasks, requiring constant improvement, and thus are always a priority. The other constant is the growing importance of the private sector to agricultural growth and the increasing importance of public sector facilitation of that growth. For agriculture to growth rapidly, good governance is critical – technically competent and committed to agricultural growth and rural development.

Technological Change

Basic science-based, institutionalized research is essential to thwart the diminishing returns incident to a limited land area, and in any case provides a high rate of return. The varied biological and physical environment of agriculture limits the transfer of technology and thus requires area-specific research systems. Because research results are often public goods, public sector research is critical to agricultural advance. As the private sector expands it will increasingly take on research activities. But even in high-income countries public sector research is a major component of private-public sector partnerships.

As farming becomes more complex and dynamic, the educational requirement of farmers increases. Concurrently, many farm children will leave agriculture for education, demanding urban jobs. Thus technology-based agricultural growth creates a strong demand-pull for increased rural education.

Because research is so important, and because it is becoming increasingly expensive, depending on expensive equipment and large coordinated teams, low- income countries must set difficult, narrow priorities for their research activities. That is one of the most important and difficult priority-setting exercises in economic development. Typically it is not done well and so research expenditure is not efficient and agricultural growth does not reach its full potential. In parallel with research are systems for the dissemination of research results. These too start heavily in the public sector and then move to a complementary mix with the private sector.

Physical Infrastructure

Agriculture's contribution to overall economic development is dependent on a steady flow of technology that requires increased inputs and produces increased output. For those processes to proceed rapidly, transaction costs must be reduced. This requires constantly upgraded roads, electrification, and telecommunications. While such physical infrastructure is naturally provided to urban areas, the dispersal of agriculture increases infrastructure costs in rural areas and makes it necessary to sequence provision geographically.

Rapid agricultural growth requires educated people in villages to provide agricultural extension, financial institutions, and modern marketing systems. Schools and clinics are of no use without trained staff. These educated people will not live in places without the full set of physical infrastructure. Thus, there is synergy between

the requirements of agriculture and the social services for physical infrastructure.

Private Sector Input Supply and Output Marketing

Rising agricultural productivity depends on massive increases in purchased input supplies as the cost of those inputs decreases and the cost of on-farm sources increases. This in turn requires rural financial markets that can mobilize national and international savings for innovating farmers and provide an outlet for farmers' savings when they reap the income benefits of improved technology.

Rising incomes and technological advance in marketing require increased quality of farm output, especially for high-value perishable commodities, and large volumes. Thus, the size and complexity of agricultural marketing increase rapidly. While family labour force-size farms preserve their competitive position in production they are at an increasing disadvantage in meeting quality and volume requirements of modern marketing systems. This challenge is best met by organizing farmers into large units for marketing purposes. This may occur through contract farming provided by large agricultural business firms, or cooperatives, or farmers' organizations. For the latter, government may play an important role in facilitating farmer organization, but must be careful not to stifle efficiency by making them in effect government institutions.

In setting their own priorities, governments must seek the means to assist the private sector in providing the input and output supply activities, and be careful not to stifle private development with onerous regulation, even while protecting consumer interests and helping to build a favorable reputation for exports.

Change Over Time in Pace and Composition of Agricultural Growth

The sources of agricultural growth change greatly over time. Yield rapidly increases in importance compared with land area. This is because of the combined effect of loss of the land frontier with population growth and exploitation and rapid increase in the efficiency of producing improved technology. The input composition switches to purchased inputs such as fertilizer and chemical pest control and off-farm marketing and processing. This rapidly increases productivity of labour and raises income.

The output composition commences with domination by cereals and root crops as the low-cost sources of calories. As incomes rise the demand for income-elastic livestock and horticultural products grows very rapidly. These are labour-intensive commodities for which physical conditions in low- and middle-income countries are usually suitable. These commodities are little restricted by land area since a modest shift of area from extensive crops allows a large increase in their production, and so the overall growth rate accelerates. An agriculture dominated by cereals is unlikely to exceed a three per cent growth rate for more than a few years. But when livestock and horticulture come to occupy over half the agricultural GDP, as happens in middle- and high-income countries, the growth rate can accelerate to between four and six per cent.

The Importance of Trade to Agricultural Growth

In low-income countries demand for agricultural products grows slowly. Consumption is largely of cereals, incomes grow slowly and demand is inelastic. At that early stage of development, agriculture has considerably greater capacity to grow than domestic markets can absorb, and achieving that growth is vital to poverty reduction and also to overall GDP growth rates. Thus, what Hla Myint (1958) referred to, as 'vent for surplus' is important to agriculture playing its role. That is to say, agricultural production must grow faster then domestic demand and the surplus exported. This drives the domestic employment multipliers as well as paying for imported capital equipment critical to overall growth.

For agricultural exports to grow a country must produce efficiently, providing constantly

improving physical infrastructure to bring down transaction costs, and constantly increasing productivity through technological change and an effective private sector capable of adapting to rapidly changing markets and constantly rising quality standards. However, these favourable policies can be nullified by unfavourable macro policy, particularly including overvalued exchange rates. Those are the most important requisites of export success. Globalization, based on declining costs of transport, facilitates access to markets, but also brings competition. Countries lagging in provision of physical infrastructure, technological change, and efficient macro policy will be losers from globalization.

Trade protection by high-income countries has been an important barrier to export success even when low- and middle-income countries become efficient and productive. Protection is particularly onerous for cotton, widely grown in quite poor countries and heavily protected and subject to export subsidies from high-income countries. Protection may also be subtle, using health rules to make it difficult for poor countries to enter high-income markets. Thus, the rate of growth of agriculture is dependent in part on negotiations to reduce both trade barriers erected by high-income countries against high-value agricultural commodities and agricultural subsidies more generally.

Foreign Aid, Agriculture and Development

Successful late starters in economic development exceed the growth rate of the frontrunners because they can catch up by drawing capital and, more important, technology and the pure science base for creating technology from their now wealthier predecessors. Foreign aid can play an important role in those transfers. This has been dramatically the case in agriculture. In Asia, the scientific base for the startling technological breakthroughs of the Green Revolution was laid by foreign aid that sponsored the key research institutions, in Mexico, then the Philippines, and finally in many other countries. These efforts were complemented by assistance to development of a host of national institutions vital to the spread of the Green Revolution and to increasing the effectiveness of agriculture ministries.

A variety of factors, including the rise of specialized lobbies that distort the distribution of foreign aid between directly productive and social activities and away from national institutions to local institutions and, most important, from national institution building to local activities, caused foreign aid to lose its effectiveness.

The late starters, particularly in Africa, were the big losers from this shift. For the late starters to achieve faster growth than their immediate predecessors will require a return to basics. A great deal has been learned about the details of agricultural growth and its contribution to overall economic development. That new information can accelerate growth beyond previous levels. But the basic principles have not changed and there must be a reversion to these if the new knowledge is to be useful. Africa and a few low-income countries in Asia and Latin await that renaissance.

See Also

- ▶ Agricultural Finance
- ▶ Agricultural Markets in Developing Countries
- ▶ Agricultural Research
- ▶ Foreign Aid
- ▶ Family Economics
- ▶ Growth and International Trade

Bibliography

Eicher, C., and J. Staatz. 1998. *International agricultural development*. 3rd ed. Baltimore: Johns Hopkins University Press.

Mellor, J. 1976. *The new economics of growth*. Ithaca: Cornell University Press.

Mellor, J. 1995. *Agriculture on the road to industrialization*. Baltimore: Johns Hopkins University Press.

Myint, H. 1958. The 'classical theory' of international trade and the underdeveloped countries. *Economic Journal* 68: 317–337.

Ravallion, M., and G. Datt. 1996. How important to India's poor is the sectoral composition of economic growth? *World Bank Economic Review* 10(1): 1–25.

Aid Conditionality

Oliver Morrissey

Abstract

Aid conditionality refers to the practice of donors attaching conditions to enhance the effectiveness of aid. The donor's prime objective is to reduce poverty, but recipients want to divert some of the aid to elites. This gives rise to two problems: adverse selection (aid does not go to the recipients who will make best use) and moral hazard (recipients can misuse the aid). The article reviews how aid conditionality can address these problems, and briefly considers empirical evidence.

Keywords

Aid agencies; Donors; IMF; Moral hazard; Structural adjustment; World Bank

JEL Classification

O190; F330

Introduction

Conditionality emerged as a major theme in the aid literature from the early 1980s with the advent of Structural Adjustment lending by the World Bank; the basic idea was not new, but the emphasis became greater. In the context of aid policy or practice, conditionality is interpreted as attaching policy reform requirements (conditions) to aid to enhance the effectiveness of the aid in promoting growth and poverty reduction. Donors believe that the policy reforms are intrinsically beneficial, so aid recipients who implement the reforms will achieve the best outcome; conditionality is good for recipients by encouraging better policies and behaviour. The latter view is contested, as briefly discussed in the final section.

The discussion here is restricted to aid and the principle of conditionality; the appropriateness of actual conditions, although an important and contentious topic, is only briefly considered in the final section. Conditions associated with debt relief or bailouts, mostly involving the IMF (or Troika in the current EU cases), are not considered; some general principles are similar, but the details and actual conditions are very different. In the aid context, the core issue is how donors can ensure that the aid they give to a recipient government is used in such a way that it benefits the poor in the country (the target group the donor wants to help). The donor's prime objective, albeit not the only one, is reducing poverty; as they believe that growth reduces poverty, conditionality in practice may relate to growth-promoting or poverty-reducing actions. Conditionality in principle simply restricts the discretion that the recipient has in using the aid.

Donors face two particular challenges because recipients differ in their willingness to comply with conditions – some will not be willing to implement the reforms. The first is avoiding adverse selection: donors want to give aid to the recipients that will make the best use of it, but may not know which recipients these are. The simple idea is that there are 'good' and 'bad' (and intermediate) types of recipients, but donors are not certain of the type. The second challenge is to avoid moral hazard, so that when recipients get aid they use it properly. An enforcement mechanism is required to ensure that bad types of recipient will implement conditions. Experience shows that such enforcement mechanisms are elusive.

Traditional conditionality, such as adjustment lending, involves allocating aid to recipients who commit to certain policy reforms so as to signal to donors that they are the good type. The aid is given first and the conditions are implemented later; Koeberle et al. (2005) refer to this as *ex ante* conditionality. The literature on credibility of policy reform highlights the limitation (Rodrik 1989; Drazen 2000): those that do not wish to reform will nevertheless commit in order to receive support, but then will not implement the reforms; genuine reformers suffer because they cannot show they are a committed type.

This form of conditionality is not effective at addressing adverse selection or moral hazard.

A strategy to avoid moral hazard is to allocate aid in tranches and monitor compliance at each stage so that the next tranche is only released if there is sufficient compliance. The threat of donors to punish recipients by not releasing a tranche in the face of non-compliance is not credible, as donors have incentives to continue to disburse aid (Mosley et al. 1991; Martens et al. 2002).

Because traditional conditionality failed to address moral hazard, donors turned to a selectivity approach where aid was given to those recipients that implemented some condition (prior action) and then received aid; Koeberle et al. (2005) refer to this as ex post conditionality. Good recipients will perceive the benefit, undertake prior actions and receive aid; bad types will not accept the cost of prior action and will reject conditional aid (Bougheas et al. 2007). Selectivity is effective to the extent that it excludes the 'worst' recipients, but this leaves donors with the problem of how to engage with 'bad' recipients with many poor people that they want to help.

The theoretical literature on aid conditionality represents this predicament with a scenario where the donors are poverty-averse and allocate aid with the objective of benefiting the poor in recipients that are less committed than the donors to reducing poverty. In this framework, aid represents a classic principal–agent problem. Conditional lending is where donors (principals) aim to design an aid contract that is attractive only to the most deserving recipients (selecting 'good' agents) or ensures an incentive for bad recipients to comply (avoiding moral hazard). The focus in the first section is on the core intuition, rather than technical detail, of the main theoretical models, while the next section considers extensions to address conditionality failures. The final section provides a brief overview of the empirical literature.

Conditionality and Aid Contracts

Theoretical papers on conditionality have three basic features to simplify reality and focus the analysis; they differ on the emphasis and formal detail. First, increasing the consumption of the poor (representing poverty reduction) is the primary objective of donors. Second, donors tend to allocate more aid to recipients with greater need (poorer countries, or countries with more poor people), which leads to adverse selection because these are more likely to be bad types. The third feature is that although recipients care about the poor, they have an incentive to limit reductions in poverty because doing so increases future expected aid inflows, creating moral hazard. This is manifested in various ways, such as corruption and weak governance (indicating bad types; Azam and Laffont 2003), so that elites capture some aid (Svensson 2000b) or exerting little effort in costly actions that would benefit the poor (Epstein and Gang 2009).

The reality that donors have to allocate aid across many recipients plays an important role. Svensson (2000a) has a model in which donors allocate aid across two recipients. Conditionality requires recipients to implement specific policies and the more effort they apply the more likely they are to be in a good state (where consumption of the poor increases). The donor allocates aid according to the state of the recipients and their reform effort. The assumption that donors have high poverty aversion has the crucial implication that they will give more aid to the recipient in the worse state (as it has more poor). If the donor commits to a pre-announced allocation and effort is observable, the poor benefit. If effort is not observable, the benefit to the poor is lower, and if additionally the donor cannot commit, the benefit to the poor is lowest. The underlying reason is that recipients have an incentive to minimise any increase in the consumption of the poor (low effort) so that they are in a relatively bad state (more poor) and therefore attract more aid than the other recipient because the donor has not pre-committed and therefore allocates according to the observed state. Conditionality only promotes effective aid under full donor commitment and observable recipient effort.

Moral hazard is driven by tensions within the recipients. Svensson (2000b) models rent-seeking between social groups over government resources. Each social group wants to capture rents for their private consumption, which reduces

the share of resources avaiable to provide public goods that benefit the poor. If social groups cooperate, the provision of public goods (welfare for society) is maximised, but rentseeking encourages non-cooperative behaviour. Aid increases government resources, but because this encourages rent-seeking the effect can be to reduce provision of public goods. The adverse effect of aid occurs because donors allocate more aid to recipients with worse outcomes (lower public goods), but if donors can commit aid, cooperation by social groups is encouraged and aid is more effective.

Retaining the rent-seeking feature, Azam and Laffont (2003) consider whether contracts that link aid to the consumption of the poor can improve on the basic income effect of aid (increased resources associated with aid allow some benefit to trickle down to the poor). The main results are that, assuming full commitment, unconditional aid does not improve on this outcome (the rich simply share the aid as additional government resources) but as long as consumption of the poor is observable the poor are better off than otherwise (aid contracts therefore increase effectiveness). If consumption of the poor is not fully observable (the donor does not know the recipient type) selectivity is required: moral hazard is avoided if the amount of aid a recipient receives depends on the level of consumption of the poor and aid is disbursed *after* observing the consumption. For good types a contract can ensure that the benefit to the poor is maintained, whereas for intermediate types the benefit is reduced, but is still better than unconditional or no aid. Aid contracts can avoid adverse selection, as there will be some bad types that receive no aid because they have no incentive to accept the contract.

One interpretation of Azam and Laffont (2003) is that an aid contract (conditionality) with full commitment is always desirable because the poor benefit (aid is effective) or the worst recipients are excluded from aid. In this way contracts mitigate the donor poverty-aversion problem that drives aid ineffectiveness in Svensson (2000a,b), even if recipient type or effort is not fully observable. However, the assumption that donors can offer complete contracts with monitoring (even if imperfect) may be unrealistic for the aid setting.

Improving Conditionality

The tendency of donors to support the poorest even if they are a bad type, because donors cannot fully commit and have no strong enforcement mechanism, gives rise to conditionality failures. In response, donors can avoid moral hazard by placing more emphasis on recipient performance and/or address adverse selection by identifying better recipients from their actions.

Avoiding Moral Hazard

Moral hazard can be avoided if donors are discouraged from allocating more aid to recipients with the greatest need. Svensson (2000a) suggests delegating aid allocation to donors with less poverty aversion, such as multilateral agencies, who will attach greater weight to recipient performance. However, Hagen (2006) shows that donors would delegate to agents with greater poverty aversion in some cases and less poverty aversion in others. Delegation is not a clear-cut solution, especially as donors like to keep control over their aid budget. Svensson (2003) offers a solution where donors commit aid to a group of recipients but subsequently disburse between those recipients according to their performance. Epstein and Gang (2009) also propose a contest between recipients where donors allocate on the basis of governance only. Recipients have to make a costly investment in improving governance, but have an incentive to do so as this determines how much aid they receive. However, such governance-based approaches give insufficient recognition to the needs of the poorest countries. Donors have to provide some aid to the poorest, even if they use aid ineffectively, but could improve aid allocation by recognising the trade-off through 'the concept of need-adjusted aid effectiveness which is a combined measure of the needs and governance quality in a country' (Bourguignon and Platteau 2012, p. 20).

Addressing Adverse Selection

Adverse selection can be addressed by using observable signals to deny aid to (bad type) recipients that attach too low a weight to consumption of the poor, so that aid is given to countries with 'a high enough quality of governance' (Azam and Laffont 2003, p. 40), on the basis that the poor have the highest consumption level in countries with good governance. However, the countries with the greatest need for aid are often those with the worst levels of governance or policy, and when aid is more plentiful (or less politically costly) donors may be less concerned about ineffective aid in 'bad' recipients as they benefit from being seen to help the poor (Bourguignon and Platteau 2012). The donors' predicament is that if they give aid where it is most needed it is likely to be least effective. This is why it can be sensible for donors to impose a cost (a prior action that increases aid effectiveness) that recipients have to bear in order to receive aid.

A Comment on Empirical Studies

There are good, or at least understandable, reasons why donors want the principle of conditionality to address moral hazard and adverse selection. However, the theoretical literature is not informative on the extent and nature of actual conditions, which are the basis of concerns about conditionality in practice. Donors, especially the World Bank (or IMF in bailouts or stabilisation), are criticised for requiring too many conditions (covering many areas of economic policy and institutional reform) and imposing reforms that are too strict (e.g. cutting subsidies or spending quickly) or optimistic (e.g. expecting rapid private sector responses to liberalisation or quick action on corruption). Extensive conditionality asks too much of recipients in too short a time. Consequently, recipients may fail to implement reforms because they are simply unable to do so and conditionality may fail to deliver expected outcomes because the conditions were inappropriate. These are legitimate concerns about the detail and practice of conditionality.

As developed from the 1980s, conditional lending required recipients to implement specified (mostly economic) policy reforms in return for being granted aid. In broad terms, recipients made progress in implementing reforms even if conditionality has not been demonstrably successful in terms of improved outcomes, whether growth or poverty reduction (Koeberle et al. 2005). There is considerable evidence that conditionality does have effects, even if weak (Koeberle 2003). Countries with adjustment programmes tend to exhibit better performance, at least in respect of social sector conditionality (Bedoya 2005) and variables under the control of donors (Malesa and Silarszky 2005). The process of conditional lending has promoted reform effort over time (Morrissey 2004); the majority of developing countries have steadily implemented reforms over the past two decades in the direction advocated by donors (for an illustration in respect of trade policy, see Jones et al. 2011).

The empirical evidence is quite limited, as it is inherently difficult to devise meaningful measures of policy reform and implementation with which to assess compliance with conditions (Morrissey 2004). Most of the evidence is based on country case studies, but too many of these are qualitative or even anecdotal. Instances where aid did not have the intended effect, or specific deficiencies in implementation are found, are cited as evidence that conditionality did not work. Conditionality may appear to fail for reasons that are beyond the control of the recipient governments (or donors). For example, shocks are more important determinants of short-term economic performance in low-income countries than aid or policy (this is implicitly recognised in Svensson (2003)). Experience does confirm the concern of theory that donors cannot use aid to leverage full implementation of conditions that recipients have no preference for, but also shows that gradual progress occurs.

After some 30 years of experience with conditional lending most donors now advocate lighter and more flexible approaches and recognise that recipients need policy space. Although few donors are clearly advocating selectivity (the US approach using governance indicators motivates

Epstein and Gang (2009), many still place some emphasis on prior actions. A flexible approach is that these actions should be negotiated in a relationship based on partnership and monitoring. Partnership implies dialogue between donors and recipients, and repeated interactions that offer ways to improve conditionality, as when donors learn more about recipients it is easier to address adverse selection and moral hazard failures.

See Also

- ▶ Development Economics
- ▶ Financial Structure and Economic Development
- ▶ Foreign Aid
- ▶ Non-governmental Organizations

Bibliography

Azam, J.-P., and J.-J. Laffont. 2003. Contracting for aid. *Journal of Development Economics* 70(1): 25–58.

Bedoya, H. 2005. Conditionality and country performance. In *Conditionality revisited: Concepts, experiences and lessons,* ed. S. Koeberle et al., pp. 187–195. Washington DC: World Bank.

Bougheas, S., I. Dasgupta, and O. Morrissey. 2007. Tough love or unconditional charity? *Oxford Economic Papers* 59(4): 561–582.

Bourguignon, F., and J.-P. Platteau. 2012. *Does aid availability affect effectiveness in reducing poverty?* Working Paper No. 2012/54. UNU-WIDER, Helsinki.

Drazen, A. 2000. *Political economy in macroeconomics*. Princeton: Princeton University Press.

Epstein, G., and I. Gang. 2009. Good governance and good aid allocation. *Journal of Development Economics* 89(1): 12–18.

Hagen, R. 2006. Samaritan agents? On the strategic delegation of aid policy. *Journal of Development Economics* 79(1): 249–263.

Jones, C., O. Morrissey, and D. Nelson. 2011. Did the World Bank drive tariff reforms in eastern Africa? *World Development* 39(3): 324–335.

Koeberle, S. 2003. Should policy-based lending still involve conditionality? *The World Bank Research Observer* 18(2): 249–273.

Koeberle, S., H. Bedoya, P. Silarszky, and G. Verheyen (eds.). 2005. *Conditionality revisited: Concepts, experiences and lessons*. Washington, DC: World Bank.

Malesa, T., and P. Silarszky. 2005. Does World Bank effort matter for success of adjustment operations? In *Conditionality revisited: Concepts, experiences and lessons,* ed. S. Koeberle et al., pp. 127–141. Washington, DC: World Bank.

Martens, B., U. Mummert, P. Murrell, and P. Seabright. 2002. *The institutional economics of foreign aid*. Cambridge: Cambridge University Press.

Morrissey, O. 2004. Conditionality and aid effectiveness re-evaluated. *The World Economy* 27(2): 153–171.

Mosley, P., Harrigan, J., and Toye, J. 1991. *Aid and power: The World Bank and policy-based lending* (two volumes). London: Routledge.

Rodrik, D. 1989. Promises, promises: Credible policy reform via signalling. *Economic Journal* 99: 756–772.

Svensson, J. 2000a. When is foreign aid policy credible? Aid dependence and conditionality. *Journal of Development Economics* 61(1): 61–84.

Svensson, J. 2000b. Foreign aid and rent-seeking. *Journal of International Economics* 51: 437–461.

Svensson, J. 2003. Why conditional aid does not work and what can be done about it? *Journal of Development Economics* 70(2): 381–402.

A

Airline Industry

Severin Borenstein and Nancy Rose

Abstract

The 1978 US airline deregulation benefited passengers through lower fares and expanded service. Airline privatization and liberalization elsewhere in the developed world has since had similar effects. Still, there have been some unanticipated effects: hub-and-spoke networks have efficiency appeal, but they also increase congestion and confer market power on dominant airlines; price discrimination is widespread; loyalty programmes exacerbate market power concerns; airline finances are subject to extreme cyclic volatility; and labour is a significant residual claimant on profits. Airline competition and industry structure remain in flux: entry and exit are commonplace, as is experimentation with new pricing and products.

Keywords

Airline deregulation; Airline industry; Market power; Price discrimination; Principal–agent conflict

JEL Classifications
L66

Since the mid-1970s, privatization and deregulation have transformed domestic passenger airline markets in many developed economies.

From its infancy through the early 1970s, scheduled passenger air service was considered a public utility nearly everywhere in the world. In most countries, this took the form of state-owned national airlines, often operating with significant government subsidies. US airlines were privately owned, but prices and entry decisions were controlled by federal regulators. California and Texas provided limited but notable exceptions, where small airlines providing only intra-state service operated free of most economic regulation. Their substantially lower fares and higher load factors relative to regulated operations foreshadowed the possible impact of deregulation.

The United States legislated federal airline deregulation in 1978, replacing government decision-making with carrier determination of pricing, entry, and network configuration. Within 20 years, similar reforms faced newly privatized and entrant carriers operating within Europe, Asia and Australia. Most international air travel, however, remains heavily regulated through bilateral government agreements, apart from intra-European Union flights and a few examples of 'open skies' pacts that allow broad freedom in entry and pricing.

Deregulation yielded numerous benefits, best documented for the US domestic market due to public availability of detailed, high-quality data. The most striking and robust finding is that fares are substantially lower and passengers are better off under deregulation than they would have been under continued regulation (in the United States) or state ownership (in many other countries); see, for example, Borenstein (1992), Morrison and Winston (1995) and Borenstein and Rose (2006). Facilitating lower prices were decreased costs per available seat–mile and increased load factors, resulting from a mix of operational reorganization, service changes, and efficiency gains. In the United States deregulation-induced transfers from labour to consumers were initially modest, though labour costs and contract negotiations have since become focal in competition between formerly regulated 'legacy' carriers and discount airline entrants in many markets. Labour transfers generally account for a more substantial share of cost reductions for newly privatized carriers.

While price declines conformed to expectations, not all responses to deregulation were anticipated. First, legacy airlines in the United States rapidly reconfigured their operations from point-to-point to hub-and-spoke networks, in which coordinated 'banks' of flights arrive at a centrally located airport, allow passengers to change planes, and depart a short time later. This allows airlines to offer relatively frequent, albeit connecting, service on a large number of city pairs without dedicating aircraft to serving each route non-stop. Legacy carriers outside the United States generally operated some form of hub-based network even prior to reform, due largely to relatively thin domestic markets and bilateral agreements that restricted international service to operate through a few gateway airports. Hub-and-spoke operations initially were thought to confer significant efficiency improvements, facilitating greater flight frequency and higher load factors for all but the most dense markets, though it was recognized that passengers preferred non-stop service, all else equal.

Over time, the benefits of hubs have been called into question. Coordinated banks of flights increase congestion costs and delays at hub airports and reduce system-wide aircraft utilization rates; airline dominance of local traffic in and out of their hubs raises concerns about market power; many hubs have been created, then abandoned, as airlines attempted to discern the optimal number and characteristics of hub airports.

Second, average real price declines masked an explosion in pricing complexity. From a pair of distance-based coach and first-class fares on each route, airlines sprouted a dozen or more fare offerings. Prices on a single carrier-route may differ by the time or day of travel, how far in advance a ticket is purchased, the length of stay, and whether the stay includes a Saturday night. Economists

have debated the extent to which fare variation reflects efficient competitive peak-load pricing or potentially less efficient price discrimination, but both effects are undoubtedly significant in most markets.

Third, market power concerns, focal at hub airports generally dominated by a single carrier, have been exacerbated by the diffusion of various loyalty programmes. Best-known are frequent flyer programmes, which reward passengers for concentrating their business with a single carrier, but similar programmes were also created for travel agents, who booked about 85 per cent of all tickets in the early deregulation days. Non-linear reward schemes benefit the largest carrier in a market and increase switching costs among their participants. These programmes also generate principal–agent conflicts: travel agents benefit from directing passengers to flights that may be slightly more expensive or less desirable in exchange for side payments from the carrier. Similarly, in exchange for free personal travel, business passengers choose flights for which their employer may have to pay more.

Fourth, extreme cyclic volatility of airline finances has raised concerns about the 'core' of the competitive equilibrium. The industry reaped large profits when demand was strong relative to capacity and fuel prices were low (the late 1980s and late 1990s) and reported huge losses when fuel prices rose and demand weakened, generating excess aircraft capacity and a wave of bankruptcies (the early 1980s, 1990s and 2000s). Debate continues over whether this profit volatility should spark concern or is part of the normal functioning of an industry with high fixed costs, slow capacity adjustment, fluctuating operating costs (particularly fuel), and highly cyclical and unpredictable demand. Is this any different from the steel, computer memory chip, or software industries which also have exhibited extreme swings? Economic research has provided few answers as yet.

Finally, airline labour has been at the heart of continuing concern and stress. At most legacy carriers, pilots and mechanics have negotiated very lucrative contracts during good times, effectively sharing in the high profits. When profits declined, however, downward adjustment of wages has been slow. Entry or expansion by new airlines with substantially lower labour pay scales is fairly easy, particularly during downturns when excess capacity makes aircraft leases cheap and easily available. During downturns, wages at established carriers may differ most from competitive wages, leaving incumbents vulnerable to new competition and financially constrained in their ability to respond aggressively. The rise of low-cost carriers and intensity of legacy carrier wage and benefit cuts in the most recent industry downturn raise significant questions for the future position of airline employees.

Many of the research results from early post-deregulation studies have been reopened in the face of dramatic industry evolution over recent years. The challenge to both economists and industry participants is to infer the long-run equilibrium structure of the industry. What is the stable number of airlines in a given geographic market? What sort of competition is feasible? Are hub networks viable in the face of point-to-point competition? What is the long-run role of labour as a quasi-equity holder? These questions remain for future researchers to address.

See Also

- ▶ Agency Problems
- ▶ Bankruptcy, Economics of
- ▶ Network Goods (Empirical Studies)
- ▶ Network Goods (Theory)
- ▶ Price Discrimination (Empirical Studies)
- ▶ Price Discrimination (Theory)

Bibliography

Borenstein, S. 1992. The evolution of US airline competition. *Journal of Economic Perspectives* 6(2): 45–73.

Borenstein, S., and Rose, N. 2006. Airline deregulation and liberalization: Lessons learned. Working paper. Cambridge, MA: NBER.

Morrison, S., and C. Winston. 1995. *The evolution of the airline industry*. Washington, DC: Brookings Institution.

Aiyagari, S. Rao (1952–1997)

Zvi Eckstein and Dan Peled

Keywords

Generational altruism; Borrowing constraints; Consumption inequality; Dynamic macroeconomics; Dynastic models; Income inequality; Idiosyncratic income shocks; Market frictions; Mobility; Numerical solution techniques; Optimal taxation; Overlapping generations models; Precautionary saving; Risk aversion; Taxation of capital; Taxation of capital income; Time preference; Uninsurable idiosyncratic risks

JEL Classifications

B31

S. Rao Aiyagari was 45 years old when he died in 1997, just as his approach to dynamic macroeconomic research was gaining recognition. Rao's vision was motivated by empirical observations and academic debates stemming from the different implications of aggregate and individual economic data. In particular, individual earnings, saving, wealth and labour exhibit much larger fluctuations over time than per-capita averages, and accordingly significant individual mobility is hidden within these cross-sectional distributions. Rao became convinced that this kind of heterogeneity and individual dynamics has important implications for the understanding of aggregate economic data and can provide new insights on the role of various economic policies.

The Aiyagari–Bewley economic model, proposed by Bewley (1986) and developed further in Aiyagari (1994, 1995), has become a leading model for modern dynamic macroeconomics. The economy is populated with heterogeneous infinitely lived agents subject to uninsurable idiosyncratic income risks. Possible long sequences of adverse income shocks naturally lead to borrowing constraints on individuals, and consequently fluctuations in consumption can be mitigated only by precautionary individual savings. Since agents' histories of income shocks are different, the model generates equilibrium cross-section distributions of wealth, saving and consumption, which reflect the fact that borrowing constraints are tighter for wealth-poor agents. These cross-sectional distributions are contrasted with or calibrated to fit their empirical counterparts in the data, and their responses to various policy changes can be analysed. Solving for the equilibrium in dynamic models with heterogeneous agents is complicated, and Rao was among the pioneers in developing and applying numerical solution techniques for that purpose.

In his most influential paper (Aiyagari 1994), Rao investigated the implications of precautionary saving due to individual earning risks and borrowing constraints for aggregate savings. He found that the contribution of uninsured idiosyncratic risks to aggregate saving is modest for plausible values of risk aversion, variability and persistence of earnings (at most three per cent), but can be significantly larger with higher variability and persistence parameters of the earning stochastic process. Access to asset markets in that model enables agents to cut consumption volatility by half, and enjoy a welfare gain of 14% of per-capita consumption compared with the equilibrium with no access to assets markets. The model generates a wealth distribution that is positively skewed, more dispersed than income distribution, and inequality is significantly higher for wealth than for income.

Precautionary savings generated by uninsured idiosyncratic shocks and borrowing constraints motivated Rao to examine the recommendation to eliminate tax on capital income (Lucas 1990). Aiyagari (1995) showed that for the Aiyagari–Bewley economies this dictum may be wrong because the frictions in these models result in agents' behaviour that is closer to that in overlapping generations (OLG) models. Precautionary saving can lead to over-accumulation of capital in equilibrium, so that positive taxes on capital are needed to bring the pre-tax return on capital to equality with the rate of time preferences, at any point in time as well as in the long run. In contrast

to OLG models, where government debt can also be used to reduce excessive saving, in Aiyagari–Bewley economies the demand for such assets becomes infinite when the interest rates approaches the rate of time preferences. The suitability of the model for addressing such fundamental issues is evidenced by the fact that a decade later it was still being used to study the same issue, albeit with different conclusions (Werning 2005).

Rao has examined many other implications of cross-sectional distributions generated by frictions in capital markets and uninsurable idiosyncratic risks, such as asset pricing and trading patterns (Aiyagari and Gertler 1991), setting taxes in a median-voter context (Aiyagari and Peled 1995), marriage patterns and investment in children (Aiyagariet al. 2000, 2002). He also studied the equilibrium implications of market frictions and borrowing constraints that emerge endogenously from private information on individual earnings (Aiyagari and Williamson 2000). Many other influential papers have adopted his framework of uninsurable idiosyncratic risks for the study of various phenomena, including, for instance, Kocherlakota (2005) on optimal taxation, Krueger and Perri (2006) on the joint evolution of income and consumption, and Storesletten et al. (2004) on age-dependent income and consumption inequality.

Rao's earlier theoretical work focused on the links between dynastic and OLG models, and provided the deep theoretical understanding of dynamic models that he applied in his subsequent work. He examined whether the two models become similar in terms of equilibrium existence, optimality and cyclicality, with and without money, when the life of each generation and the period of overlap across generations are sufficiently long, or when generations are linked through altruism (for example, 1985; 1988; 1989). Additional work with Wallace and others examined the role for policy in search equilibrium models of money (for example, Aiyagari et al. 1996; Aiyagari and Wallace 1997).

Aiyagari published more than 30 influential papers during his 18-year career as an economist. The force of his work and ideas and their impact on his colleagues are evidenced by the continued appearance of his co-authored papers for many years after his unexpected death, exhibiting some of the most innovative dynamic macroeconomic research.

A

See Also

- ▶ Income Taxation and Optimal Policies
- ▶ Incomplete Markets

Selected Works

1985. Observational equivalence of the overlapping generations and the discounted dynamic programming frameworks for one-sector growth. *Journal of Economic Theory* 35: 202–21.

1988. Nonmonetary steady states in stationary overlapping generations models with long lived agents and discounting: multiplicity, optimality, and consumption smoothing. *Journal of Economic Theory* 45:102–27.

1989. Can there be short-period deterministic cycles when people are long lived? *Quarterly Journal of Economics* 104: 163–85.

1991. (With M. Gertler.) Asset returns with transaction costs and uninsured individual risks. *Journal of Monetary Economics* 27: 311–31.

1994. Uninsured idiosyncratic risks and aggregate saving. *Quarterly Journal of Economics* 109: 659–84.

1995. Optimal capital taxation with incomplete markets, borrowing constraints, and constant discounting. *Journal of Political Economy* 103: 1158–75.

1995. (With D. Peled.) Social insurance and taxation under sequential majority voting and utilitarian regimes. *Journal of Economic Dynamics and Control* 19: 1511–28.

1996. (With N. Wallace and R. Wright.) Coexistence of money and interest-bearing securities. *Journal of Monetary Economics* 37: 397–419.

1997. (With N. Wallace.) Government transaction policy, the medium of exchange, and welfare. *Journal of Economic Theory* 74: 1–18.

2000. (With J. Greenwood and N. Guner.) On the state of the union. *Journal of Political Economy* 108: 213–44.

2000. (With S. Williamson.) Money and dynamic credit arrangements with private information. *Journal of Economic Theory* 91: 248–79.
2002. (With J. Greenwood and S. Ananth.) Efficient investment in children. *Journal of Economic Theory* 102: 290–321.

Bibliography

Bewley, T. 1986. Stationary monetary equilibrium with a continuum of independently fluctuating consumers. In *Contributions to mathematical economics in Honor of Gérard Debreu*, ed. W. Hildenbrand and A. Mas-Colell. Amsterdam: Elsevier Science Pub. Co.
Kocherlakota, N. 2005. Zero expected wealth taxes: A Mirrlees approach to dynamic optimal taxation. *Econometrica* 73: 1587–1621.
Krueger, D., and F. Perri. 2006. Does income inequality lead to consumption inequality? Evidence and theory. *Review of Economic Studies* 73: 163–193.
Lucas Jr., R. 1990. Supply-side economics: An analytical review. *Oxford Economic Papers* 42: 293–316.
Storesletten, K., C. Telmer, and A. Yaron. 2004. Consumption and risk sharing over the life cycle. *Journal of Monetary Economics* 51: 609–633.
Werning, I. 2005. Tax smoothing with redistribution. Staff Report No. 365, Federal Reserve Bank of Minneapolis.

Akerlof, George Arthur (Born 1940)

Brian G. M. Main

Abstract

George Akerlof is forever associated with his landmark 1970 paper, 'The market for "lemons"', which transformed the way economists approach markets where there is a difference between the transacting agents in the information they possess. This concept of asymmetric information, with its major impact on many fields of economics, was singled out when, in 2001, he was awarded the Nobel Memorial Prize in Economics (along with Michael Spence and Joseph Stiglitz). A more comprehensive assessment of his contribution to economics would be as providing a better behavioural underpinning for macroeconomics as a major figure in the New Keynesian movement.

Keywords

Akerlof, G; Asymmetric information; Caste system; Efficiency wage theory; Friedman, M; Neuroeconomics; New Keynesian economics; Inflation-unemployment trade-off; Market for lemons; Social norms

JEL Classification

B31; E; E0

George Akerlof's father came to the United States from Sweden to obtain a Ph.D. at the University of Pennsylvania, and remained in the country to pursue a career as a research chemist. He met George's mother while she was a graduate student in chemistry. Hers was an academic family. George's great grandfather was among the earliest graduates from the University of California at Berkeley (in 1873), and his grandfather also graduated from Berkeley. Other members on that side of the family also established successful academic careers. George grew up on the East Coast, where his father held a series of posts, variously at Yale University, at the Mellon Institute in Pittsburgh and at Princeton University, before running his own independent research firm in the Princeton area. Indeed, it was witnessing the uncertainty surrounding his father's continuing employment, dependent as it was on securing government research grants, which first turned George Akerlof's mind to macroeconomic themes such as unemployment. As an undergraduate at Yale he majored in mathematics and economics, and in the fall of 1962 he entered graduate school at MIT, where he had the good fortune to find himself one of an exceptionally talented cohort of students. His doctoral supervisor was Robert Solow (Nobel Laureate 1987). Akerlof joined the Berkeley faculty in the fall of 1966 and, although he has spent extended periods away from Berkeley – at the Indian Statistical Institute in New Delhi, the Council of Economic Advisors, the Federal Reserve Board (where he met his wife, Janet Yellen), the LSE, and the Brookings

Institution – he has remained closely identified with Berkeley ever since.

The 'Market for "Lemons"' Paper

For the generations of economics students trained since 1970, when asked to single out a favorite economics article, it is a pretty safe bet that the most popular article would be George Akerlof's (1970) paper on asymmetric information, 'The market for "lemons"'. Part of this paper's appeal lies in its modelling approach. While mathematically rigorous, it is derived from close observation of the world. Care is taken to incorporate realistic economic detail, yet the results obtained provide tremendously powerful insights. The reader is left with an understanding of an important market situation that was previously obscure and, in addition, is offered policy options whereby economic well-being can be improved. This general approach characterizes all of Akerlof's work.

The 'lemons' paper starts by offering an analysis of the second-hand car market in which the existence of lower-quality vehicles (the eponymous 'lemons') can disrupt the workings of the market – to the extent that the usual economic law of lowering the price in the face of an excess of supply (or difficulty experienced in selling into the market) simply makes matters worse. Rather than bringing about a market equilibrium through matching supply and demand, the lower price drives out the better-quality cars remaining in the market and this further depresses demand.

The problem arises from an asymmetry of information that exists between those supplying used cars into the market (they know, in considerable detail, just how good or otherwise their present car is) and those who are buying in the market (they can obviously inspect the car, but are left with substantially less knowledge than the seller). If those on the demand side use the price as an indication of the average quality of car traded, this can cause demand to decline in the face of falling prices – if, as seems reasonable, the suppliers with better-quality cars withhold them as the price falls, leaving only the poorer-quality cars to be offered at lower prices. Note that this problem does not arise in the new car market. While this market is, unfortunately, not free from 'lemons', the probability of being stuck with a lemon can be ascertained from sources such as consumer reports. The fraction of new cars entering the market as lemons does not vary with the price or discount offered on new cars.

Varian (1992, p. 469) offers the following simple characterization of the model. Assume there is a quality-of-car index q, which is uniformly distributed between 0 and 1. Additionally, assume the demand for cars is a function of this quality to the extent that the price offered for cars of quality q is exactly $(3/2)q$ and that, on the other side of the market, suppliers with a car of quality q would be willing to sell for price q or better. There is clearly scope for mutually beneficial trade in this market, as any price between q and $(3/2)q$ leaves both the buyer and seller of a car with quality q better off.

On the other hand, if the buyer is unable to perceive the quality of the car but has to rely on the average quality of cars traded in the second-hand market as a measure of the expected quality of any car purchased, then the price offered is $(3/2)\ q^*$, where q^* is the average quality in the market.

But on the supply side, of course, sellers know the exact quality of their cars and, for any price p, only those with quality p or lower will offer cars for sale. Thus, the observed quality of cars traded at price p will be $p/2$. However, at quality $p/2$ there will be no cars demanded, as cars of this average quality fetch an offer of only $(3/2)\ q^* = (3/2)(p/2) = (3/4)p$. So no cars will traded at this price. But nor will a fall in the price offer any improvement because, if price falls, then so too will the quality of car offered to the market and the average quality of cars observed. As things stand, there is no price that will allow cars to be traded. Potentially mutually advantageous trades are not made. Economic welfare is lower than it might be. The culprit is, of course, asymmetric information.

It is the inability of the supply side of the market (which possesses the hidden information about car quality) to meaningfully communicate this information to the buyers that undermines the potential for mutually advantageous trades.

The existence of lemons inhibits the proper functioning of the market. Akerlof points out that the inability of older people to secure health-care insurance, the inability of minorities to secure decent employment prospects, the external costs of dishonest business practices, and the difficulty developing countries experience in establishing capital markets can all be viewed as manifestations of the same 'lemons' problem, i.e., asymmetric information.

In awarding the 2001 Nobel Memorial Prize in Economics to George Akerlof, Michael Spence and Joseph Stiglitz, the Royal Swedish Academy of Sciences cited 'their analyses of markets with asymmetric information'. In reviewing the contributions of these prize winners, Rosser (2003) identifies a nascent discussion of this idea in the earlier economics literature, but there is little doubt that it was with the publication of Akerlof's 1970 'Market for "lemons"' paper that the metaphorical light bulb was switched on in the economics community and the idea of asymmetric information started to become integrated into economics. As a recent survey by Riley (2001) makes clear, this concept is now an important feature of modern approaches to development economics, financial economics, industrial organization, international economics, labour economics, and many other areas. It is now difficult to imagine the world of economics without this insight.

Other Work

While for many people the 'lemons' paper stands as a seminal example of the power of microeconomic analysis, the underlying motivation that led Akerlof to investigate this area was actually macroeconomic. Cyclical fluctuations in the car market were seen as a major destabilizing factor in the macroeconomy: hence the original research effort. Throughout his career Akerlof has been driven by a desire to develop macroeconomics in a way that allows problems such as unemployment to be better understood. Never happy with the neoclassical synthesis and distinctly critical of the New Classical economics, Akerlof has been a major contributor to the development of New Keynesian Economics (2002). Indeed, his work can be seen as a lifetime effort to create a better behavioural microfoundation to macroeconomics – continuing in the tradition started by Keynes' (1936) *General Theory.*

Caste and Identities

In subsequent work the 'lemons' paper was soon developed into an analysis of caste systems (1976, 1985), in which irrational and economically inefficient belief systems can be sustained out of a concern for individual well-being, albeit at the cost of society's overall welfare. This work is typical of Akerlof's approach to economic theory in that it seeks to broaden our view of economic exchange from the simplistic dyad of buyer and seller (the focus of so much economic analysis) to admit the real possibility that such exchanges are heavily conditioned by the existence of wider social forces. In this specific case, people adhere to what are obviously dysfunctional behaviours because, in their individual calculus, the costs of being seen to break such conventions (and hence being outcaste) outweigh any individual short-term gains. Thus, individually rational action leads to a macroeconomically inefficient outcome.

More generally, people can be seen as exhibiting patterns of behaviour that are consistent with chosen identities but would be otherwise difficult to explain (Akerlof and Kranton 2000). Such identities are chosen in an attempt to fit most comfortably into society, given people's individual circumstances. The choice of identity brings with it a set of behaviours and an exposure to the behaviour of others with whom one identifies. This stream of work represents a major step in bridging the gap between economics and sociology that is so aptly summarized by James Duesenberry (quoted in Granovetter 1985, p. 485): 'economics is all about how people make choices; sociology is all about how they don't have any choices to make.'

This approach led Akerlof to empirical analyses of the dramatic rise in out-of-wedlock births (Akerlof et al. 1996) and the marked increase in the number of men living without children (1998). These papers demonstrate that the rise of children born to unmarried mothers and the increase in men living outside of households with children can each be ascribed to changing norms

(the notion of the shotgun marriage and the destigmatization of out-of-wedlock births) that have more to do with changing technology (birth control) and the social reaction to these changes than to any wealth or incentive effects arising from welfare programmes.

This enthusiasm to engage with real-world data and empirical work is another salient characteristic of Akerlof's work. Somewhat unusually, for a theorist of major repute, he has throughout his career undertaken empirical studies of the major social and economic policy issues of the day. Thus, in addition to the analysis of family structure and poverty mentioned above, he has studied the distribution of employment and unemployment experience (Akerlof and Main 1980, 1981), job mobility (Akerlof et al. 1988), German reunification (Akerlof et al. 1991), financial malfeasance (Akerlof and Romer 1993), and the inflation-unemployment trade-off (Akerlof et al. 1996, 2000). Akerlof's intellectually open and outgoing approach to his work also shows in the wide range of co-authors involved in his theoretical work, including, for example, Akerlof and Miyazaki (1980), Akerlof and Milbourne (1980), Akerlof and Katz (1989), Akerlof and Yellen (1990), and Akerlof and Kranton (2005). As will be seen below, his collaboration with Janet Yellen has been the most sustained and intellectually productive.

Near-Rational Economic Behaviour

While the 'lemons' paper is undoubtedly his most famous, the stream of papers that best demonstrates Akerlof's New Keynesian pedigree starts with Akerlof (1969). This paper investigates structural unemployment in a framework that sees firms as being in monopolistic competition and having staggered price setting, with wages emerging as bargains struck between firms and workers. With Taylor's (1979) incorporation of rational expectations, this links directly to the overlapping contracts approach that now lies at the heart of the New Keynesian model. Akerlof also deployed this approach in the study of monetary policy (1973, 1978, 1979). Here, simple monitoring rules by agents of their bank balances are shown to make both monetary and fiscal policy effective.

Extending this approach more generally, Akerlof and Yellen (1985) demonstrate that what appear as rule-of-thumb behavioural rules deployed in economic decision-making actually bring with them substantial savings in computational costs (and deal with the bounded rationality problem) while, at the same time, imposing only second-order costs on the agent by way of lost economic efficiency. In this sense, such rules of thumb are quite sustainable and sensible modes of behaviour. The insights of this paper have far-reaching implications. Accepting the existence of such behaviour not only points to why monetary policy might be effective but also explains why there can, indeed, be significant trade-offs between inflation and unemployment, particularly at low rates of inflation (Akerlof et al. 1996, 2000).

Friedman's (1968) original attack on the notion of a long-run trade-off between inflation and unemployment was further strengthened by the incorporation of rational expectations by the New Classical economists, Lucas (1972) and Sargent (1971). Deploying the Akerlof and Yellen (1985) insight of near-rational behaviour towards inflation, Akerlof et al. (2000) demonstrate that at low rates of inflation, such as were typical in the 1950s and are now prevalent once again, there can be an empirically significant trade-off between inflation and unemployment. The fact is that in setting wages and prices economic agents (business people, wage negotiators and so on) do not behave exactly as economic models of rational expectations would suggest – at least not when inflation is moderate and the costs of deviating from such rationality are modest when compared with the informational and computational costs involved.

Sociologically Based Efficiency Wage Theory

In attempts to explain the unemployment that fiscal and monetary policy is often deployed to remedy, a standard question is why in the face of unemployment wages do not simply decline, so restoring equilibrium in the market. The answer is, of course, that cheaper is not always better. In a paper evocatively titled 'Jobs as dam sites', Akerlof (1981) explains that, just as it makes poor economic sense to construct a lower-quality dam on a prime site (no matter that it may be cheaper), so it may not

make economic sense to hire cheaper labour even when available. These ideas, further developed in Akerlof (1982) and most elegantly expressed in Akerlof and Yellen (1990), provide a sociologically rooted explanation for efficiency wages.

The key idea here is that the exchange between employer and employee is rich and complex, extending well beyond the narrow instrumental delivery of labour in return for wages. Workers who display 'consummate' cooperation in playing their part to achieve the objectives of the organization are much preferred to those exhibiting 'perfunctory' cooperation (see Williamson et al. 1975, p. 266). Part of the key to ensuring the higher-productivity outcome is being seen to pay a fair wage. The concept of fair wage-effort is socially determined, and both equity theory from social psychology and social exchange theory from sociology offer explanations of how workers react when this balance is disturbed. From this perspective, the financial savings from lowering wages can be a poor bargain when set against the impact on the productivity of the workforce. In the face of such rigidity coming about through the individually rational decisions of employers, there is clear scope for macroeconomic policy to effect a coordinated move to a higher level of employment. This is a key insight of the efficiency wage model of the labour market (Akerlof and Yellen 1986).

Psychologically Based Models

The incorporation of psychological insights into economics has proved highly successful in recent years, as indicated by the award of the Nobel Prize in 2002 to Daniel Kahneman. Akerlof and Dickens (1982) is an early contribution to this movement, drawing on the notion of cognitive dissonance whereby individuals choose their beliefs or view of a situation in such a way that renders them the greatest comfort or happiness. In this way, it is possible to explain many common phenomena that otherwise seem to make little economic sense, such as the widespread flouting of workplace safety standards. In some ways the more recent work in Akerlof and Kranton (2005) on choice of identity can be seen as a sociological version of this same phenomenon. The common theme is that social actors are capable of choosing the frame through which they view their circumstances and, unsurprisingly, can be expected to choose an approach that, given the situation in which they find themselves, offers them the greatest comfort. To an external observer this can often result in behaviours that are perplexing.

Thus, in Akerlof (1991) a psychologically based explanation is offered for the widely documented phenomenon of people acting in ways that seem too short-sighted to be in their interest. This is seen in the widespread failure to make adequate provision for retirement or to save enough in general. Drawing on a personal experience during a year living in India during the late 1960s, Akerlof recounts how day after day he procrastinated over mailing off a promised package to Joseph Stiglitz. This is developed into a model that demonstrates why in repeatedly opting for what appears as the best short-term course of action (to procrastinate) one is often left in a situation that in retrospect one may regret. The insights offered by this model of economic behaviour are both powerful and far-reaching, and later proponents, such as David Laibson (1997), have extended the area into neurological studies of the brain under the heading 'neuroeconomics'.

Conclusion

If economists were ever to adapt the psychologist's stimulus-response technique into a game of declaring a famous economist's name as a stimulus and then noting the response, it seems clear that the overwhelming response to 'George Akerlof' would be 'lemons'. This would, at the same time, be both a sufficient response and an insufficient response. As the above discussion has shown, it is insufficient to try to capture such a major body of important studies by reference to one paper. Akerlof has not only dealt with asymmetric information but, as a major contributor to modern Keynesian economics, has also confronted the major macroeconomic issues of the day, most notably by providing the

behavioural underpinnings to explain the efficacy of interventionist economic policy.

Yet the 'lemons' response could arguably be judged sufficient in the sense that the 'lemons' paper contains all of the elements that make Akerlof's approach to economic theory so different and so potent. Mark Granovetter (1985) criticizes economic models as either totally ignoring the influence of social structures and relations or else going to the other extreme, by being oversocialized in the sense that there are really no choices left for agents to make. Akerlof is one of a small but growing set of economists who manage to position their models on the middle ground. Far from Friedman's (1953) positive economics approach, which regards assumptions as something to be minimized and whose realism is of no consequence as long as the predictive power of the model holds up, Akerlof adheres to an approach that utilizes models based on closely observed empirical examples. The fact that the most observers believe that monopolistic competition is the norm means to Akerlof that such a feature must appear in the model. A model utilizing perfect competition might be able to do just as well, but would be rejected in the face of Akerlof's pragmatic goal of making his models as near to the observed reality as possible while still being tractable.

'The market for "lemons"' will almost certainly stand as Akerlof's best-known contribution, having provided the impetus for radical new ways of looking at events in so many areas of economics. But it is also an excellent exemplar of a different approach to economic modelling. It is this pragmatic approach to economic modelling that makes all of Akerlof's contributions so worthwhile.

See Also

- ▶ Caste System
- ▶ Economic Sociology
- ▶ Efficiency Wages
- ▶ Information Aggregation and Prices
- ▶ Social Norms

Selected Works

1969. Relative wages and the rate of inflation. *Quarterly Journal of Economics* 83, 353–374.

1970. The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84, 488–500.

1973. The demand for money: A general-equilibrium inventory-theoretic approach. *Review of Economic Studies* 40, 115–130.

1976. The economics of caste and of the rat race and other woeful tales. *Quarterly Journal of Economics* 90, 599–617.

1978. The microfoundations of a flow of funds theory of the demand for money. *Journal of Economic Theory* 18, 190–215.

1979. Irving Fisher on his head: The consequences of constant target-threshold monitoring for the demand for money. *Quarterly Journal of Economics* 93, 169–187.

1980 (With B. Main.) Unemployment durations and unemployment experience. *American Economic Review* 70, 885–893.

1980. (With R. Milbourne.) Irving Fisher on his head II: The consequences of the timing of payments for the demand for money. *Quarterly Journal of Economics* 94, 145–157.

1980. (With H. Miyazaki.) The implicit contract theory of unemployment meets the wage bill argument. *Review of Economics Studies* 47, 321–338.

1981. Jobs as dam sites. *Review of Economic Studies* 48, 37–49.

1981. (With B. Main.) An experience-weighted measure of employment and unemployment durations. *American Economic Review* 71, 1003–1011.

1982. Labor contracts as partial gift exchange. *Quarterly Journal of Economics* 97, 543–569.

1982. (With W. Dickens.) The economic consequences of cognitive dissonance. *American Economic Review* 72, 307–319.

1985. Discriminatory status-based wages among tradition-oriented stochastically based coconut producers. *Journal of Political Economy* 93, 265–276.

1985. (With J. Yellen.) A near-rational model of the business cycle, with wage and price inertia. *Quarterly Journal of Economics* 100, 823–838.
1986. (With J. Yellen, eds.). *Efficiency wage models of the labor market*. New York: Cambridge University Press.
1988. (With A. Rose and J. Yellen.). Job switching and job satisfaction in the US labor market. *Brookings Papers on Economic Activity*, 2, 495–592.
1989. (With L. Katz.) Workers' trust funds and the logic of wage profiles. *Quarterly Journal of Economics* 103, 525–536.
1990. (With J. Yellen.) The fair-wage effort hypothesis and unemployment. *Quarterly Journal of Economics* 105, 255–283.
1991. Procrastination and obedience. *American Economic Review* 81(2), 1–19.
1991. (With A. Rose, J. Yellen and H. Hessenius.) East Germany in from the cold: The economic aftermath of currency union. *Brookings Papers on Economic Activity*, 1, 1–105.
1993. (With P. Romer.) Looting: The economic underworld of bankruptcy for profit. *Brookings Papers on Economic Activity*, 2, 1–60.
1996. (With W. Dickens and G. Perry.) The macroeconomics of low inflation. *Brookings Papers on Economic Activity* 1, 1–59.
1996. (With J. Yellen and M. Katz.) An analysis of out-of-wedlock childbearing in the United States. *Quarterly Journal of Economics* 111, 277–317.
1998. Men without children. *Economic Journal* 108, 287–309.
2000. (With W. Dickens and G. Perry.) Near-rational wage and price setting and the optimal rates of inflation and unemployment. *Brookings Papers on Economic Activity* 2000: 1, 1–60.
2002. Behavioral macroeconomics and macroeconomic behavior. *American Economic Review* 92, 365–394.
2000. (With R. Kranton.) Economics and identity. *Quarterly Journal of Economics* 115, 715–753.
2005. (With R. Kranton.) Identity and the economics of organizations. *Journal of Economic Perspectives* 19, 9–32.

Bibliography

Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*, ed. M. Friedman. Chicago: University of Chicago Press.
Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58: 1–17.
Granovetter, M. 1985. Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* 91: 481–510.
Keynes, J. 1936. *The general theory of employment, interest and money.* London: Macmilan.
Laibson, D. 1997. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics* 112: 443–477.
Lucas Jr., R. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.
Riley, J. 2001. Silver signals, twenty-five years of screening and signaling. *Journal of Economic Literature* 39: 432–478.
Rosser Jr., J. 2003. A Nobel Prize for Asymmetric Information: The economic contributions of George Akerlof, Michael Spence and Joseph Stiglitz. *Review of Political Economy* 15: 3–21.
Sargent, T. 1971. A note on the 'accelerationist' controversy. *Journal of Money, Credit, and Banking* 3: 721–725.
Taylor, J. 1979. Staggered wage setting in a macro model. *American Economic Review* 69: 108–113.
Varian, H. 1992. *Microeconomic analysis*, 3rd ed. New York: W. W. Norton & Company.
Williamson, O., M. Wachter, and J. Harris. 1975. Understanding the employment relationship: The analysis of idiosyncratic exchange. *Bell Journal of Economics* 6: 250–278.

Åkerman, Johan Gustav (1888–1959)

K. Velupillai

Gustav Åkerman received perhaps the supreme accolade for any economist working in the theory of capital, Knut Wicksell's endorsement. Wicksell concluded his masterly review of the first part of Åkerman's doctoral dissertation with the following acknowledgement: 'I am convinced that on the whole the author has made a really significant contribution to the theory of capital' (Wicksell 1934, Appendix 2(a), p. 273).

Born in 1888, Åkerman obtained his doctoral degree from the law faculty of the University of Lund in the days before economics as a subject had independent status, in 1923. He was appointed Docent (Associate Professor) in Lund the same year, on the strength of his brilliant doctoral dissertation, 'Realkapital und Kapitalzins'. He was subsequently appointed Professor of Political Economy and Sociology at what was later to become the University of Gothenburg in 1931, and remained there until his retirement. He died in 1959.

Wicksell's famous two-part review article (the second part being on 'Åkerman's Problem') of the first volume of his dissertation assured him international fame. The first volume of his dissertation dealt with the static problems of fixed-capital systems and the second volume with dynamic problems for analogous systems. His method of analysis, in the Austrian tradition, was very similar to Böhm-Bawerk's approach: copious numerical and special examples to illustrate subtle and deep general propositions. It is to his great credit that he seldom went wrong in deriving propositions by this primitive method; as a testimony to his insights we can cite concepts and issues at the frontiers of capital theoretic debates that owe much to the results of his dissertation of 1923–1924: Wicksell effects, truncation of production flows, transverse flows, to name but a few.

He was perhaps also the first (after the early classical economists) to try to approach the problem of fixed capital as joint products – a method made famous by von Neumann and Sraffa in more recent times.

Even before his capital theoretic writings, he had engaged the grand old man of Swedish economics, Knut Wicksell, in a debate in the pages of the *Ekonomik Tidskrift* (1922) on the latter's proposals on norms for price stabilization.

His later work was mostly on practical problems of economic policy.

Selected Works

1923. *Realkapital und Kapitalzins: Heft* I. Stockholm: Centraltryckeriet.

1924. *Realkapital und Kapitalzins: Heft* II. Stockholm: Centraltryckeriet.

1931. Om den Industriella Rationaliseringen och dess Verkningar (bilaga 2 till Arbetslöshetsutredningens Betänkande). *SOU* 42, Stockholm.

1947. *Engelsk Arbetslöshet och Arbetslöshetspolitik*. Göteborg.

1956. Marginal productivity with different agricultural products. In *Twenty five economic essays in honour of Erik Lindahl*. Stockholm: Ekonomisk Tidskrift.

Bibliography

Wicksell, K. 1934. Real capital and interest: Dr. Gustav Åkerman's *Realkapital und Kapitalzins*. Appendix 2(a) of K. Wicksell, *Lectures on Political Economy.* London: Routledge & Kegan Paul.

Åkerman, Johan Henrik (1896–1982)

K. Velupillai

Somewhat lesser known internationally than his elder brother Gustaf, Johan Henrik Åkerman was, however, much better known inside Sweden. He was a prolific contributor to the theoretical, methodological, epistemological and policy debates in Sweden for almost 50 years. He challenged almost single-handedly (at least inside Sweden) the methodological position of the so-called Stockholm School and made valiant (but unsuccessful) attempts to provide an alternative vision which he described as the 'Lund School' method.

Johan Henrik Åkerman was born in Stockholm in 1896, graduated from the Stockholm Business School in 1918, and then spent two terms at Harvard University (1919–20) working with Warren M. Persons. On his return to Sweden, he continued his postgraduate studies in the Universities of Uppsala and Sweden. He obtained his PhD (Fil.Dr.) in 1929 from the University of Lund, where he was appointed Associate Professor in Political Economy

and Economic Statistics in 1932. In 1943 he was appointed Professor in Political Economy in Lund, and retained that position until his retirement in 1961. His scientific publications of more than 150 items included several books, some of which were translated into English and German. He was almost totally deaf from a very early stage in his life and totally deaf during his tenure as Professor in Lund. He died in 1982.

Johan Åkerman's outstanding doctoral dissertation had the title *On the Rhythm of Economic Life* (*Om Det Ekonomiska Livets Rytmik*). It was an ambitious attempt to codify, theoretically and empirically, all aspects of the problem of fluctuations in economic life. It was based on the theoretical framework of Wicksell's *Geldzins und Gűterpreise* (1898) and Cassel's 'Om kriser och dåliga tider' (On crises and bad times), *Ekonomisk Tidskrift*, 1904; and on the empirical methodology of the budding NBER work. Åkerman's dissertation was perhaps the earliest attempt to apply spectral analysis for studying time series phenomena in economics. His main examiner for the doctoral degree was Ragnar Frisch, whose more influential later work on 'Propagation problems and impulse problems in dynamic economics' (1933) owes much to Åkerman's specific considerations of Wicksell's celebrated 'rocking-horse' example. This latter example, delineating one influential strand in business cycle methodology – the stochastic approach – stressed the important distinction between sources of propagation and impulse mechanisms. It is to Åkerman's great credit that he was able to revive and place in the centre of discussion on business-cycle methodology this important distinction, which was initially stressed by Wicksell in an obscure footnote to a review article in the *Ekonomisk Tidskrift*, 1918. It is both important and topical in view of recent developments in equilibrium business cycle theories, where these issues are central. Indeed, Åkerman's dissertation could claim to be an early manifesto of aspects of the New Classical Economics.

In the 1930s and 1940s Åkerman's research interests shifted towards methodological and epistemological problems – mainly under the influence and impact of the works of members of the Stockholm School (and later the Keynesians). He was severely critical of the rationality and individualistic assumptions underlying the then popular macroeconomic theories (and their microeconomic underpinnings). He developed a highly original alternative modelling strategy for macroeconomics based on a so-called dual principle of 'causal' and 'computing' ('Kalkyl') models where institutional details and socio-economic classes were explicit factors. His research and reflections on these matters, spread over a period of 30 years, were summarized and elegantly delivered as a lecture on the occasion of his retirement ('Avskedsföreläsning') from the Professorship in Lund on 9 May 1961 ('Fyra methodologiska moment', *Ekonomisk Tidskrift*, 1961). The depth of his understanding of recent developments in economic analysis, and the scope of his comprehensive references to epistemological developments in theoretical physics and relevance to economic theory, were displayed in that last masterly lecture.

His lifelong interest in the political economy of business cycles was also reflected in a highly original work on political business cycles, *Ekonomiskt Skeende och Politiska Förändringar*. He was continuing a Swedish tradition on this subject – and quite independently of Kalecki's important work on political business cycles – initiated by Herbert Tingsten's inter-war work on *Political Behaviour: Studies in Election Statistics* (1937).

Retrospectively, it is significant that Johan Åkerman's two pioneering studies on problems of fluctuations in mixed economies have their counterpart in research in the frontiers of the theory and empirical analysis of business, political and economic cycles even today.

Selected Works

1928. *Om Det Ekonomiska Livets Rytmik*. Stockholm: Nordiska Bokhandeln.

1932. *Industriförbundets Produktionsindex: Motiv och Principer*. Stockholm: Sveriges Industriförbund.

1934. *Konjunkturteoretiska Problem*. Lund: C.W.K. Gleerup.
1936. *Ekonomisk Kausalitet*. Lund.
1939. *Ekonomisk Teori, I: De Ekonomiska Kalkylerna*. Lund/Leipzig.
1944. *Ekonomisk Teori, II: Kausalanalys av det Ekonomiska Skeendet*. Lund/Leipzig.
1945. *Banbrytare och Fulföljare inom Nationalekonomien*. Lund: C.W.K. Gleerup.
1946. *Ekonomiskt Skeende och Politiska Förändringar*. Lund: C.W.K. Gleerup.
1952. Innovationer och Kumulativa Förlopp. *Ekonomiskt Tidskrift*.
1960. Samhällsstruktur och Ekonomisk Teori. *Samhällsvetenskapliga studier* 18 (Lund).
1961. Fyra Metodologiska Moment. *Ekonomiskt Tidskrift*.

Bibliography

Frisch, R. 1933. Propagation problems and impulse problems in economic dynamics. In *Economic essays in honour of Gustav Cassel*. London: Allen & Unwin.
Tingsten, H. 1937. *Political behaviour: Studies in election statistics*, Stockholm Economic Studies, No. 7. London: P.S. King & Son.

Albert the Great, Saint Albertus Magnus (c.1200–1280)

Odd Langholm

Keywords
Albert the Great; Aquinas, St. T; Aristotle; Just price; Labour; Private property; Usury; Value

JEL Classifications
B31

Albert the Great, *doctor universalis*, was the foremost German philosopher and theologian of the Middle Ages. He was born in the village of Lauingen on the Danube and became a member of the Dominican Order while studying at Padua. He subsequently studied at Paris, and eventually taught there as well as in Dominican houses in Germany, primarily Cologne, where he became Regent Master of Studies and where he died. Albert served as Bishop of Regensburg, was German Provincial of his Order and Master of the Sacred Palace of the Pope, but repeatedly returned to Cologne to devote himself to study and teaching. He composed a comprehensive set of commentaries on the works of Aristotle and is considered the founder of Christian Aristotelianism. He was canonized and named a Doctor of the Church in 1931. Ten years later he was declared patron 'of all who cultivate the natural sciences', which indicates his main area of interest. In what is now called economics he is overshadowed by his famous student Thomas Aquinas, but in fact he made important contributions of his own. They are found in his comments on Scripture and on the theological *Sentences* of Peter Lombard as well as in some of his Aristotelian works. On the *Nicomachean Ethics* he composed a close textual commentary, and later a freer *Ethica.* His *Politica* is the first complete Latin commentary on Aristotle's *Politics.*

Two striking features of Albert the Great's discussions of matters relating to material wealth and economic activity are his empirical orientation and the store he sets by human labour. He argues that private property is the best arrangement in civil society because common ownership engenders strife, pointing to the observable fact that those who reap less than their labour share under communism are likely to protest and cause trouble *(Politica,* II.2). In Book V of the *Nicomachean Ethics,* Aristotle discusses justice in relation to barter between persons of different occupations and states obscurely that as one person is to another person, thus are their respective products to each other. Albert the Great interprets this formula in terms of respective input: as a farmer is to a shoemaker in labour and expenses, thus the product of the shoemaker is to the farmer's product *(Ethica,* V.2.9). This solution is explained by a factual observation: unless a carpenter receives for a bed what it cost him to make it, he will not make any more beds *(Ethica,* V.2.7).

In his commentary on the *Sentences,* Albert's approach and conclusion are different. In the absence of economic coercion and fraud, the just price is that at which a good sold can be valued according to the estimation of the market at the time of the sale *(Comm. Sent.,* IV.16.46). If these arguments are combined, what Albert asserts is that the competitive market determines value but that unprofitable goods will be withdrawn from the market.

Albert discussed the purposes and properties of money and warns against debasement of the currency. Examining usury in the same context, he rejects the 'barren metal' theory falsely attributed to Aristotle. Lending for profit is a perverse use of money, which makes it *seem as though* money reproduces itself *(Politica,* I.7). Usury is a form of economic coercion because it is paid with a conditional, not an absolute, will. The payment is voluntary only in the sense in which, according to Aristotle, the captain of a ship in peril jettisons cargo voluntarily *(Comm. Sent.,* III.37.13). But the full force of Albert the Great's denunciation of usury comes through in one of his Gospel commentaries: 'By hard labour [the borrower] has acquired something on which he could live, and this the usurer, suffering no distress, spending no labour, fearing no loss of capital by misfortune, takes away, and through the distress and labour and changing luck of his neighbour collects and acquires riches for himself' *(Super Lucam,* 6.35).

See Also

- ▶ Aquinas, St Thomas (1225–1274)
- ▶ Just Price
- ▶ Scholastic Economics

Selected Works

Opera Omnia, ed. A. Borgnet. Paris: Louis Vives.
Ethica, vol. 7.
Politica, vol. 8.
Super Lucam, vols. 22–3.
Super Tertium Sententiarum, vol. 28.
Super Quartum Sententiarum, vols 29–30.

Bibliography

Langholm, O. 1992. *Economics in the medieval schools*. Leiden: Brill.

Weisheipl, J.A. 1980. The life and works of St Albert the Great. In *Albertus Magnus and the sciences: Commemorative studies 1980*, ed. J.A. Weisheipl. Toronto: Pontifical Institute of Mediaeval Studies.

Alchian, Armen Albert (born 1914)

Steven N. S. Cheung

Alchian was born in Fresno, California, in 1914. During his economic education at Stanford he inherited from his statistics teacher, Allen Wallis, an insatiable curiosity about real-world observations. Alchian invariably aims toward the derivation of testable implications. Whether his subject is charity, tenure, organization, money, inflation, or unemployment, he adheres firmly to the elementary principles of price theory.

His professional recognition began in 1950 with the publication of his paper on evolution and economic theory. This became an instant classic which launched a vigorous debate on economic methodology destined to enliven more than a decade. The work argues that the postulate of maximization may be false but that its use is justified by the tenets of 'survival of the fittest' under competition.

Justly famous, but not widely adopted because of its radical departure from traditional cost curves, is Alchian's seminal work on cost and output, published in 1959. Here he submits that in any productive activity the faster the production rate, the higher will be the unit cost because of diminishing returns, whereas the larger the production volume, the lower will be the unit cost because of greater choice in production methods. All costs, both average and marginal, are stated in discounted present values and expressed in terms of varying production programmes. Acceptance of this approach requires substantial modification of standard supply/demand analysis.

In all likelihood, Alchian will be best remembered for his works on property rights. To him, the economic system in any society is defined by its property rights, which constitute the 'rules' of competition. When these rights are altered, competitive behaviour will change, along with changes in income distribution and resource allocation. The use of price as a criterion for competition is inherent with private property rights and, in Alchian's view, it is less important to understand how price is determined than to understand what price does as a criterion for individuals competing for economic goods.

The Alchian approach to analyse property rights in terms of pricing and competition complements R.H. Coase's approach in terms of delimitation and enforcement of rights. Far from conflicting, the two merge in powerful accord. Together, they form a core in modern economic analysis, where the paradigm of property rights has now been firmly nailed in place.

Selected Works

1950. Uncertainty, evolution and economic theory. *Journal of Political Economy* 58: 211–21.
1953. The meaning of utility measurement. *American Economic Review* 43: 26–50.
1959. Costs and outputs. In *The allocation of economic resources*, ed. M. Abramovitz. Stanford: Stanford University press.
1962a. (With R.A. Kessel.) Competition, monopoly and the pursuit of money. *Aspects of Labour Economics*, pp. 157–75.
1965b. (With R.A. Kessel.) Effects of inflation. *Journal of Political Economy*, 70: 521–37.
1965. Some economics of property rights. *Il Politico* 30, 816–29.
1968. Cost. In *International Encyclopedia of the Social Sciences*. New York: Macmillan, Vol. 3. 404–15.
1969. Information costs, pricing and resource unemployment. *Economic Inquiry* 7: 109–28.
1972. (With H. Demestz.) Production, information costs and economic organization. *American Economic Review* 62: 777–95.
1977. *Economic Forces at Work*. Indianapolis Liberty Press.
1978. (With B. Klein and G.C. Crawford.) Vertical integration, appropriable rents and the competitive contracting process. *Journal of Law and Economics* 21: 297–326.
1983. (With W.R. Allen). *Exchange and Production*. Belmont: Wadsworth.

Algeria, Economy of

Barry Turner

Keywords
Arab Spring; Dinar; International Monetary Fund

JEL Classification
O53; R11

Overview

The economy is heavily dependent on public spending and the sale of oil and gas. In 1994 an IMF-sponsored programme for economic reconstruction reduced inflation and set in place a free-market economy. However, it was a further decade before privatization made significant strides.

The economy has achieved growth every year since 1995, with rising oil prices prompting robust growth from 2003. Algeria was largely shielded from the global financial crisis thanks to its limited exposure to international financial markets and a very low external debt. However, with hydrocarbons accounting for 98% of exports, there is a need for economic diversification. The services and construction sectors are gradually expanding and non-hydrocarbon growth averaged 6% per year over the decade from the early 2000s.

The creation of new private sector jobs is essential as unemployment remains high despite growth and government initiatives bringing it

down to 10% in 2010 from 27% in 2000. Unemployment is especially acute among women and the young. While Algeria faced relatively little disruption during the Arab Spring, regional fragility threatens longer-term growth prospects.

In 2009 petroleum and natural gas (excluding refined petroleum) contributed 31.0% to GDP; followed by transport, communications, trade, restaurants, finance, real estate and services, 25.1%; public administration and defence, 10.9%; public utilities and construction, 10.9%.

Currency

The unit of currency is the *Algerian dinar* (DZD) of 100 *centimes*. Foreign exchange reserves were US$146,130 m. in September 2009, with gold reserves 5.58 m. troy oz. Total money supply was 4,071.5 bn. dinars in June 2009. Inflation rates (based on IMF statistics):

2002	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*
1.4%	2.6%	3.6%	1.6%	2.3%	3.6%	4.9%	5.7%	3.9%	4.5%

The dinar was devalued by 40% in April 1994.

Budget

The fiscal year starts on 1 January. In 2009 budgetary central government revenue totalled 3,740,500 m. dinars and expenditure 2,556,900 m. dinars. Principal sources of revenue in 2009 were: taxes on income, profits and capital gains, 2,231,700 m. dinars; taxes on goods and services, 1,048,900 m. dinars. Main items of expenditure by economic type in 2009: compensation of employees, 860,500 m. dinars; social benefits, 555,600 m. dinars.

VAT is 17%.

Performance

Real GDP growth rates (based on IMF statistics):

2002	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*
4.7%	6.9%	5.2%	5.1%	2.0%	3.0%	2.4%	2.4%	3.3%	2.4%

Total GDP was US$205.8 bn. in 2012.

Banking and Finance

The central bank and bank of issue is the Banque d'Algérie. The *Governor* is Mohammed Laksaci. In 2002 it had total reserves of US$23.5 bn. Private banking recommenced in Sept. 1995. In 2002 there were five state-owned commercial banks, four development banks, nine private banks and two foreign banks.

Foreign debt fell from US$25,388 m. in 2000 to US$16,871m. in 2005 and further to US$5,276 m. in 2010 (representing just 3.4% of GNI).

There is a stock exchange in Algiers.

La Banque d'Algérie: http://www.bank-of-algeria.dz

See Also

- Energy Economics
- International Monetary Fund
- Islamic Economic Institutions
- Islamic Finance
- Oil and the Macroeconomy
- Organization of the Petroleum Exporting Countries (OPEC)

Alienation

Jonathan Wolff

Abstract

The term 'alienation' is associated especially with the early writings of Karl Marx, for whom the core idea was that of human beings becoming detached from part of their 'essence'. At one stage Marx hoped to demonstrate that all of the concepts of classical economics could be derived from the concept of alienation. As he matured, exploitation and surplus value replaced alienation at the heart of his analysis. Nevertheless, Marx's observations regarding alienation remain insightful to this day.

Keywords
Alienation; Capitalism; Classical economics; Communism; Engels, F.; Exploitation; Feuerbach, L.; Hegel, G. W. F.; Historical materialism; Marx, K. H.; Smith, A.; Surplus value

JEL Classifications
B00

Although the word 'alienation' is commonly used to express an idea of, perhaps, resentful dislocation, within social theory its central use is to be found in the early writings of Karl Marx (1818–83), and especially his *Economic and Philosophical Manuscripts,* also known as the *Paris Manuscripts,* of 1844. Marx did not invent the concept; it was widely used by the group of Young Hegelian philosophers with whom he associated in the early 1840s, and especially by Ludwig Feuerbach, in his account of religious alienation. In turn, these thinkers had been influenced by Hegel's concept of externalization.

The term itself cannot be given a single, uncontroversial definition; rather. it seems a marker for a constellation of ideas, not always present in every use. A common understanding sees alienation as a subjective feeling. For Marx, however, alienation is an objective fact about the world, and in its core use we can often distinguish three constitutive elements. The most easily observable aspect is that human beings become detached from something that properly belongs to them. This implies, of course, a second element; a normative claim about how things ought to be, that is, their non-alienated state. Finally, and most metaphysically ambitious, that from which man has become separated nevertheless returns in some 'alien' form; by this means human beings are not only estranged from but also dominated by their own essence or products.

Marx's use of the idea of alienation went through a number of phases. The first takes over and extends Feuerbach's concept of religious alienation. The second is the most ambitious: alienation is used as an explanatory concept in the sense that it is claimed that all the categories of economics can be generated from an analysis of the concept of alienation. This neo-Hegelian phase, however, was short-lived, not surviving beyond the *Economic and Philosophical Manuscripts*; Marx was shortly to become aware that a priori philosophy was not the best tool for economic analysis. In a third phase the idea of alienation was retained as a central concept in the understanding of the effect of capitalism on human beings, and held out the promise of emancipation. This, however, faded in to a fourth and final phase where, although, the same ideas were present, the term itself was used less and less, and Marx's key concept for the analysis of capitalism became surplus value or exploitation.

Religious Alienation: The Influence of Feuerbach

The young Marx wrote for a philosophical audience which had accepted Feuerbach's reversal of traditional theology in which he asserted that human beings had created God in their own image; indeed this is a view with a long history. Feuerbach's distinctive contribution was to argue that worshipping God diverted human beings from enjoying their own human powers. While accepting much of Feuerbach's account, Marx criticized Feuerbach on the grounds that he had failed to understand why people fell into religious alienation and so was unable to explain how it could be transcended. Marx's explanation, of course, was that religion was a response to alienation in material life, and could not be removed until human material life was emancipated, at which point religion would wither away. This was discussed in Marx's 1843 essay *Contribution to the Critique of Hegel's Philosophy of Right*: *Introduction,* and, very briefly, in the *Theses on Feuerbach* of 1845.

Precisely what it is about material life that creates religion was not set out by Marx with complete clarity. However, it seems that at least two aspects of alienation are responsible. One is alienated labour, which will be explored shortly. A second is the need for human beings to assert their communal essence. Marx argued that, whether or not we explicitly recognize it, human beings exist as a community, and what makes

human life possible is our mutual dependence on the vast network of social and economic relations which engulf us all, even though this is rarely acknowledged in our day-to-day life. Marx's view appeared to be that we must, somehow or other, acknowledge our communal existence in our institutions. At first it is 'deviously acknowledged' by religion, which creates a false idea of a community in which we are all equal in the eyes of God. After the post-Reformation fragmentation of religion, when religion is no longer able to play the role even of a fake community of equals, the state fills this need by offering us the illusion of a community of citizens, all equal in the eyes of the law. But the state and religion will both be transcended when a genuine community of social and economic equals is created.

Here we see all three aspects of alienation. Human communal existence has come apart from its essence through the invention of God. The normatively correct situation for humans, however, is one in which they enjoy their essence on earth. Finally, our own communal essence returns to dominate us in the alien form first of religion and then of the political state.

Alienated Labour as the Foundation of Economics: The Neo-Hegelian Phase

It is commonplace to observe that Marx transformed a critique of religion into a critique of society. The *Economic and Philosophical Manuscripts* is an important element in this early critique. Here Marx famously depicts workers under capitalism as suffering from four types of alienated labour. First, they are alienated from their products, in at least two ways: they may not understand what they are making, and, as soon as it is created, is taken away from them. Second, they are alienated in productive activity (work) which is experienced as a torment, often requiring the performance of mindless or back-breaking toil. Third, they are alienated from their species-being. The distinctive feature of human beings is their productive and creative power. Yet under capitalism humans produce blindly and not in accordance with their truly human powers. Consequently, argues Marx, workers feel free only when away from work, engaged in activities that they share with animals; eating, drinking and having sex. Hence they are alienated from their distinctively human powers. Finally, they are alienated from other human beings, where the relation of exchange replaces mutual need.

These categories overlap in some respects, but this is no surprise given Marx's remarkable methodological ambition in these writings. Essentially he attempted to apply a Hegelian deduction of categories to economics, trying to demonstrate that all the categories of bourgeois economics – wages, rent, exchange, profit, and so on – were ultimately derived from an analysis of the concept of alienation. Consequently, each category of alienated labour was supposed to be deducible from the previous one. However, Marx got no further than a rather unconvincingly attempt to deduce categories of alienated labour from each other. Quite possibly in the course of writing he came to understand that a different methodology was required for approaching economic issues. Nevertheless, we are left with a very rich text on the nature of alienated labour.

Alienation and Emancipation

Marx based his account of capitalism not, at this stage, on independent empirical study, but on his readings of the works of the classical economists, most notably Adam Smith; much of the descriptive content of the idea of alienated labour from was derived his reading of *The Wealth of Nations*. However, by setting it within the theory of alienation he was able to depict capitalism as a world which was by its nature contrary to the human essence, and therefore with an inbuilt tendency to its own destruction.

The bridge between Marx's early analysis of alienation and his later social theory is the idea that the alienated individual is 'a plaything of alien forces', albeit alien forces which are themselves a product of human action. In our daily lives we take decisions that have unintended consequences, which then combine to create large-scale social forces which may have an utterly

unpredicted effect. In Marx's view the institutions of capitalism – themselves the consequences of human behaviour – come back to structure our future behaviour, determining the possibilities of our action. For example, for as long as a capitalist intends to stay in business he must exploit his workers to the legal limit. Whether racked by guilt or not, the capitalist must act as a ruthless exploiter. Similarly, the worker must take the best job on offer; there is simply no other sane option. But by doing this we reinforce the very structures that oppress us. The urge to transcend this condition, and to take collective control of our destiny – whatever that would mean in practice – was one of the motivating and sustaining elements of Marx's attraction to communism.

However, Marx's idea of emancipation – of a non-alienated society – has largely to be inferred from its negative. There are, however, two short passages in the early writings which are often cited in this context. The more famous is from the *German Ideology,* co-authored with Engels in 1845, and like many of their works unpublished in their lifetime. Here Marx and Engels described future society as a rural idyll, lived in complete freedom to order one's own life. Recent scholarship, however, casts doubt on whether this passage, which is quite unlike anything else written by Marx and Engels, was intended as a serious contribution to the development of their view (Carver 1998).

A second short passage appears at the end of the text 'On James Mill' (1844) in which non-alienated labour is briefly described in terms which emphasize both the producer's immediate enjoyment of production as a confirmation of his or her powers, and the idea that production is to meet the needs of others, thus confirming for all parties our human essence as mutual dependence. Both sides of our species essence are revealed here: our individual human powers and our membership in the human community.

Alienation and the Rise of 'Surplus Value'

As Marx turned to economics he found philosophy of decreasing use and interest, and as he matured as a social thinker the concept of alienation becomes less and less prominent. This has led some commentators, notably Althusser, to argue that there was an 'epistemological break' between Marx's early, humanist, phase, and a later scientific phase, incorporating the first volume of *Capital* (1867). Although the publication, since Althusser's famous essay (Althusser 1970), of many of Marx's writings of the 1850s shows that there is something closer to a natural development of ideas rather than a decisive break, it is true that the concept of alienation does not play the central role in Marx's later economic writings that it did in his early writings. Nevertheless, even in *Capital* there are descriptions of the labour process under capitalism which bear close comparison with the arguments of the 1844 manuscripts, and a discussion of 'commodity fetishism' in *Capital* is very close indeed to the idea of alienation.

Conclusion

Although Marx's economic theories play little role in contemporary economic analysis, and his theory of historical materialism is valued more for its small-scale insights rather than its long-term predictions, Marx's theory of alienation remains of great interest. On a descriptive level, Marx's account of the conditions of work under capitalism remain highly relevant if not to the developed world, then clearly to the major developing economies. Furthermore, the idea that human beings can become trapped within structures they have created for themselves is a deep insight that is constantly being rediscovered especially within the feminist and environmental movements. Marx's ideas concerning alienation are an inspiration even to those who are unaware of their source.

See Also

- ▶ Braverman, Harry (1920–1976)
- ▶ Capitalism
- ▶ Commodity Fetishism
- ▶ Engels, Friedrich (1820–1895)

- ▶ Exploitation
- ▶ Marx, Karl Heinrich (1818–1883)
- ▶ Socialism

Bibliography

All the writings of Marx and Engels discussed are available in McLellan (2000) and online at http://www.marxists.org. Accessed 4 Nov 2006.
Althusser, L. 1970. *For Marx*. London: Verso, 2005.
Avineri, S. 1970. *The social and political thought of Karl Marx*. Cambridge: Cambridge University Press.
Braverman, H. 1975. *Labor and monopoly capital*. New York: Monthly Review Press.
Carver, T., ed. 1991. *The Cambridge companion to Marx*. Cambridge: Cambridge University Press.
Carver, T. 1998. *The post-modern Marx*. Manchester: Manchester University Press.
Colletti, L., ed. 1992. *Karl Marx: Early writings*. London: Penguin.
Elster, J. 1985. *Making sense of Marx*. Cambridge: Cambridge University Press.
Feuerbach, L. 1957. *The essence of Christianity*. New York: Harper.
Hook, S. 1950. *From Hegel to Marx*. New York: Humanities Press.
Kolakowski, L. 1978. *Main currents of Marxism*. Vol. 1. Oxford: Oxford University Press.
Maguire, J. 1972. *Marx's Paris writings*. Dublin: Gill and Macmillan.
McLellan, D. 1970. *Marx before Marxism*. London: Macmillan.
McLellan, D., ed. 2000. *Karl Marx: Selected writings*. 2nd ed. Oxford: Oxford University Press.
Meszaros, I. 1975. *Marx's theory of alienation*. London: Merlin Press.
Ollman, B. 1971. *Alienation: Marx's conception of man in capitalist society*. Cambridge: Cambridge University Press.
Wolff, J. 2002. *Why read Marx today?* Oxford: Oxford University Press.
Wood, A. 1981. *Karl Marx*. London: Routledge.

Allais Paradox

Maurice Allais

Abstract

The 'Allais paradox' is that risk-averse persons' choices between alternatives tend to vary according to the absolute amounts of potential gain involved in different pairs of alternatives, even though rational choice between alternatives should depend only on how the alternatives differ. But there is no paradox once we accept the non-identity of monetary and psychological values and the importance of the distribution of cardinal utility about its average value.

Keywords

Allais paradox; Allais, M.; Bernoulli, N.; Cardinal utility; de Finetti, B.; Expectations; Harsanyi, J. C.; Independence; Preferences; Propensity for risk; Psychological versus monetary value; Random choice; Psychology of risk; Samuelson, P. A.; Savage, L. J.; St Petersburg paradox; von Neumann and Morgenstern

JEL Classifications

D81

The St Petersburg Paradox and the Bernoullian Formulation

Let there be a random prospect g_1, $\cdots$, g_i, $\cdots$, g_n, $\cdots$, p_1, $\cdots$, p_i, $\cdots$, p_n ($\sum_i p_i = 1$) giving the probability p_i of positive or negative gains g_i. The early theorists of games of chance considered that a game was advantageous when the mathematical expectation

$$M = \sum_i p_i g_i \quad (1 \leq i \leq n) \tag{1}$$

The principle of the mathematical expectation of monetary gains has proven to be open to question in the case of the *St Petersburg Paradox* outlined by Nicolas Bernoulli. For this game, we have: $g_i = 2^i$, $p_i = 1/2^i$, $n = \infty$ so that $M = \infty$. However, if the unit of value is the dollar, it can be seen that for most subjects, the psychological monetary value of the game (that is the price they are ready to pay for this random

prospect) is generally lower than 20 dollars. This, at first sight, involves a paradox.

To explain this paradox, Daniel Bernoulli (1738) considered the mathematical expectation of cardinal utilities $u(C + g_i)$ instead of the mathematical expectation of monetary gains, C being the player's capital. Thus the formulation (1) is replaced by the Bernoullian formulation

$$u(C + V) = \sum_i p_i u(\mathrm{C} + \mathrm{g}_i) \qquad (2)$$

in which V is the psychological monetary value of the random prospect. He proposed to take the logarithmic expression $u = \log(C + g)$ as cardinal utility (Bernoulli 1738; Allais 1952b, p. 68; 1977, pp. 498–506; 1983, p. 33). It can then be shown that we have approximately $V \sim a + [\mathrm{logC}/\log 2]$ with $a = 0.942$, which yields $V \sim 14$ or 18 US $ for C equal to 10,000 or 100,000 dollars respectively (Allais 1977, p. 572).

The Neo-Bernoullian Formulation

In order to measure cardinal utility from random choice, von Neumann and Morgenstern demonstrated in the *Theory of Gamevs* (1947), on the basis of a set of more or less appealing postulates, the existence of an index $B(C + g)$, such that

$$B(C + V) = \sum_i p_i B(C + g_i) \qquad (3)$$

in which the index $B(C + g)$ is independent of the random prospect considered, but depends on the subject (von Neumann and Morgenstern 1947, pp. 8–31 and 617–32; Allais 1952b, p. 74; 1977, pp. 521–3, 591–603; 1983, p. 34).

Using other sets of postulates, Marschak, Friedman and Savage, Samuelson, Savage, etc. (Marschak 1950, 1951; Friedman and Savage 1948; Samuelson 1952; Savage 1952, 1954; Allais 1952b, pp. 74–5, 88–92, and 99–103; 1977, pp. 464–5, 508–14; 1983, pp. 33–5) came to the same formulation (3), which may be referred to as the neo-Bernoullian formulation, but its interpretation differs depending on the postulates adopted. While von Neumann and Morgenstern believed, at least initially, that $B \equiv u$, the p_i being objective probabilities (Allais 1952b, p. 74; 1977, pp. 591–2), Savage held that cardinal utility is a myth (Savage 1954, p. 94), and that the neo-Bernoullian index B alone is real, the p_i being subjective probabilities, the existence of the function B and the p_i being proven on the basis of the axioms considered. Some authors (e.g. de Finetti, Krelle, Harsanyi) admit the existence of cardinal utility u, but they consider that $B \neq u$, and the index B is deemed to take account of the relative propensity for risk corresponding to the distribution of cardinal utility (de Finetti 1977; Allais 1952b, pp. 123–4; 1983, pp. 30–31).

Whereas von Neumann's and Morgenstern's opinion, accepted by most authors, is that the crucial axiom of their theory is axiom 3 Cb, I consider that their axioms 3 Ba and 3 Bb are the crucial ones (Allais 1977, pp. 596–8). However, one way or another, irrespective of the nature of the axioms from which it is derived, the neo-Bernoullian formulation boils down to assuming the independence of the B_i for given values of the p_i. This is the principle of independence (Allais 1952b, pp. 88–90 and 98–9; 1977, pp. 466–7).

The Allais Paradox

When I read the *Theory of Games* in 1948, formulation (3) appeared to me to be totally incompatible with the conclusions I had reached in 1936 attempting to define a reasonable strategy for a repetitive game with a positive mathematical expectation (Allais 1977, pp. 445–6). Consequently, I viewed the principle of independence as incompatible with the preference for security in the neighbourhood of certainty shown by every subject and which is reflected by the elimination of all strategies implying a non-negligible probability of ruin, and by a preference for security in the neighbourhood of certainty when dealing with sums that are large in relation to the subject's capital (Allais 1952b, pp. 84–6, 88–90, 92–5; 1977, pp. 451, 466–7, 491–8).

This led me to devise some counter-examples. One of them, formulated in 1952, has become famous as the 'Allais Paradox'. Today, it is as widespread as its real meaning is generally misunderstood.

This counter-example consists of two questions, the gains considered being expressed in (1952) francs [one million (1952) francs is roughly equivalent to 10,000 (1985) dollars].

Do you prefer Situation A to Situation B?
Situation A
certainty of receiving 100 million.
Situation B
a 10 per cent chance of winning 500 million,
an 89 per cent chance of winning 100 million,
a 1 per cent chance of winning nothing.
Do you prefer Situation C to Situation D?
Situation C
an 11 per cent chance of winning 100 million,
an 89 per cent change of winning nothing.
Situation D
a 10 per cent chance of winning 500 million,
a 90 per cent chance of winning nothing.

It can be shown that, according to the neo-Bernoullian formulation, the preference $A > B$ should entail the preference $C > D$, and conversely (Allais 1952b, pp. 88–90; 1977, pp. 533–41).

However, it is observed that for very careful persons, well aware of the probability calculus and considered as rational, and whose capital C is relatively low by comparison with the gains considered, the preference $A > B$ can be observed in parallel to the preference $C < D$. Since the neo-Bernoullians consider the axioms from which they deduce the neo-Bernoullian formulation as evident, they consider this result a paradox.

In 1952, Savage's answers to these two questions contradicted his own axioms. The explanation he gave is somewhat surprising. It boiled down to stating: 'Since my axioms are totally evident, my answers, which are indeed incompatible with my axioms, are explained by the fact that I did not give the matter enough thought' (Savage 1954, pp. 101–103).

Empirical Research

After analysing the answers to the 1952 Questionnaire (Allais 1952d). I found that the rate of violation of the neo-Bernoullian formulation corresponding to the Allais Paradox was approximately 53 per cent (Allais 1977, p. 474).

This violation example is not an isolated one (Allais 1977, pp. 636–6, n. 15). There is even one test for which the rate of violation is 100 per cent. It is based on the comparative analysis of, on the one hand, the monetary value x' attributed to a probability of 1/2 of winning a sum between 0.0001 and 1000 million, with a probability of 1/2 of winning nothing at all; and, on the other hand, of the monetary value x'' attributed to a probability p_i between 0.25 and 0.999 of winning 200 million, with a probability $1 - p_i$ of winning nothing at all. The two indexes $B1/2$ and $B200$ deduced from these two series of questions, which according to the neo- Bernoullian formulation should be totally identical up to a linear transformation, in fact are completely different for all the subjects who answered the questions. Such was in particular the case of de Finetti (Allais 1977, pp. 612–13, 620–31; 1983; pp. 61–2 and 110–11, n. 146).

Much empirical research has been carried out since 1952. It has shown that many subjects who can be viewed as rational may behave in contradiction with the neo-Bernoullian formulation (e.g. MacCrimmon and Larsson 1975; Allais 1977, pp. 507–8, pp. 611–54). Confronted with these results, the neo-Bernoullians always explain these violations as 'anomalies', 'errors', 'insufficient thought by the subjects', or 'ill constructed and inconclusive' experiments made by incompetent persons, 'inexperienced in experimental psychology' (e.g. Amihud 1974 and 1977; Morgenstern 1976). But these statements do not hold in the face of the very numerous violations observed by the many researchers, following different methods and operating in different countries at different times (Allais 1977, pp. 541–2; 1983, p. 66).

The Allais Paradox, a Simple Illustration of Allais's General Theory of Random Choice

These violations can be explained very simply. Limiting consideration to the mathematical expectation of the B_i involves neglecting the basic element characterizing psychology vis-à-vis risk, namely the distribution of cardinal utility about its mathematical expectation (Allais 1952b, pp. 51–5, 96–7; 1977, pp. 481–2, 520–23, 550–52; 1983, pp. 30–31), and in particular, when very large sums are involved in comparison with the psychological capital of the subject, the strong dependence between the different eventualities (g_i, p_i), and the very strong preference for security in the neighbourhood of certainty.

My 1952 inquiry (Allais 1952d, 1977, pp. 447–9, 451–4, 604–54; 1983, pp. 28 and 41) showed that all the subjects questioned were able to answer questions on the intensity of their preferences for different possible gains, setting aside any consideration of random choices (only a few neo-Bernoullian authors refused to answer these questions) (Allais 1943, pp. 156–77; 1952b, pp. 43–6; 1977, pp. 460–61, 475–80, 614–17, 632–3). The analysis of the answers made it possible to design a well defined cardinal utility curve, the structure of which is the same for all the subjects up to a linear transformation. It portrays their answers on average remarkably well (Allais 1984a, c).

This result is all the more significant in that this expression of cardinal utility shows a very striking similarity to the expression for psychophysiological sensation as a function of luminous stimulus, determined by Weber's and Fechner's successors (Allais 1984c, § 4.3 and Charts III and XXV).

The existence of a cardinal utility $u(C+g)$ being proven and the neo- Bernoullian index $B(C+g)$, if it exists, being defined also up to a linear transformation, it can be shown that the two indexes are necessarily identical up to a linear transformation (Allais 1952b, pp. 97–8, 103, 128–30; 1977; pp. 465, 483, 604–607; 1983, pp. 29–30; 1985).

As a consequence the neo-Bernoullian formulation reduces to considering the mathematical expectation of cardinal utility alone, neglecting its dispersion about the average. In so doing, it neglects what may be considered as the specific element of risk (Allais 1952b, pp. 49–56; 1983; pp. 35–41).

In fact the cardinal utility corresponding to a monetary value V of a random prospect should be considered as a function

$$u(C+V) = F[u(C+g_1), \ldots, u(C+g_i), \ldots \ldots, u(C+g_n), p_1, \ldots, p_i, \ldots, p_n] \quad (4)$$

of cardinal utilities u_i corresponding to the different gains g_i. Since utilities u_i are defined up to a linear transformation, it can be shown that (Allais 1977, pp. 481–3, 550–52, 607–609; 1985, § 12 and 22)

$$u + \Delta = F(u_1 + \Delta, \ldots, u_i + \Delta, \ldots\ldots, u_n + \Delta, p_1, \ldots, p_i, \ldots, p_n) \quad (5)$$

in which Δ is any constant (property of cardinal isovariation). Consequently it can be shown that relation (4) can be written

$$u(C + \ V) = \overline{u} \ + R(\mu_2, \ldots, \mu_l, \ldots, \mu_{2n-1}) \quad (6)$$

in which $\overline{u}$ represents the mathematical expectation of the u_i and the μ_l represent the moments of order l:

$$\mu_l = \sum_i p_i (u_i - \overline{u})^l. \quad (7)$$

The ratio can $\rho = R/\overline{u}$ be considered as an index of the propensity for risk. For the $\rho = 0$, behaviour is Bernoullian; for $\rho > 0$, there is a propensity for risk; for $\rho < 0$, there is a propensity for security. For a given subject, ρ can be nil, positive or negative, depending on the domain of

the field of random choices considered (Allais 1983, pp. 35–41; 1985).

The mistake made by the proponents of the neo-Bernoullian formulation is to want to impose restrictions on the preference index

$$I = f[g_1, \quad \ldots, g_i \quad \ldots, g_n, p_1, \ldots, p_i, \ldots, p_n] \quad (8)$$

of any subject other than those corresponding to conditions of rationality, such as the existence of a field of ordered random choice or the axiom of absolute preference. According to this axiom, taking two random prospects g_i, p_i and g_i', p_i such that $g_i > g_i'$ for any p_i, the first is obviously preferable to the second (Allais 1952b, pp. 38–41; 1977, pp. 457–8, 530–35; 1985, § 31.3).

Imposing other restrictions would, in the case of certain goods (A),(B),. . .,(C), reduce to imposing special restrictions on the preference index $I(A,B,\ldots,C)$ which no author has ever envisaged. In fact, to have a marked preference for security in the neighbourhood of certainty together with a preference for risk far from certainty is not more irrational than preferring roast beef to chicken (Allais 1952b, pp. 65–7; 1977, pp. 527–33; 1983, pp. 39–40; 1985, § 31.3).

From the St Petersburg Paradox to the Allais Paradox

In sum, just as the St Petersburg Paradox led Daniel Bernoulli to replace the principle of maximization of the mathematical expectation of monetary values by the Bernoullian principle of maximization of cardinal utilities, the Allais Paradox leads to adding to the Bernoullian formulation a specific term characterizing the propensity to risk which takes account of the distribution as a whole of cardinal utility (Allais 1978, pp. 4–7; 1977, pp. 548–52; 1983, pp. 35–42).

Neither the St Petersburg nor the Allais Paradox involves a paradox. Both correspond to basic psychological realities: the non-identity of monetary and psychological values and the importance of the distribution of cardinal utility about its average value.

For nearly forty years the supporters of the neo-Bernoullian formulation have exerted a dogmatic and intolerant, powerful and tyrannical domination over the academic world; only in very recent years has a growing reaction begun to appear. This is not the first example of the opposition of the 'establishments' of any kind to scientific progress, nor will it be the last (Allais 1977, pp. 518–46; 1983, pp. 69–71, 112–14).

The Allais Paradox does not reduce to a mere counter-example of purely anectodal value based on errors of judgement as too many authors seem to think without referring to the general theory of random choice which underlies it. It is fundamentally an illustration of the need to take account not only of the mathematical expectation of cardinal utility, but also of its distribution as a whole about its average, basic elements characterizing the psychology of risk.

See Also

► Expected Utility Hypothesis

Bibliography

Allais, M. 1943. *A la recherche d'une discipline économique*, Première partie: l'économie pure. Ateliers Industria, 920 pp. Second edition under the title *Traité d'économie pur* Paris: Imprimerie Nationale, 1952, 5 vols. (The second edition is identical to the first, apart from the addition of a new introduction, 63 pp).

Allais, M. 1952a. Fondements d'une théorie positive des choix comportant un risque et critique des postulats et axiomes de l'école Américaine. International Conference on Risk, Centre National de la Recherche Scientifique, May 1952. *Colloques Internationaux XL, Econométrie*, Paris, 1953, 257–332.

Allais, M. 1952b. The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American school. English translation of 1952a. In Allais and Hagen (1979), 27–145.

Allais, M. 1952c. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica* 21(4) (1953): 503–546. This paper corresponds to some parts of Allais, 1952a.

Allais, M. 1952d. La psychologie de l'homme rationnel devant le risque – la théorie et l'expérience. *Journal de la Société de Statistique de Paris*, January–March 1953: 47–73.

Allais, M. 1977. The so-called Allais' Paradox and rational decisions under uncertainty. In Allais and Hagen (1979), 437–699.
Allais, M. 1978. Editorial introduction, foreword. In Allais and Hagen (1979), 3–11.
Allais, M. 1983. The foundations of the theory of utility and risk. In *Progress in decision theory*, ed. O. Hagen and F. Wenstop, 3–131. Dordrecht: Reidel, 1984.
Allais, M. 1984a. L'utilité cardinale et sa détermination – hypothèses, méthodes et résultats empiriques. Memoir presented to the Second international conference on foundations of utility and risk theory, Venice, 5–9 June 1984.
Allais, M. 1984b. The cardinal utility and its determination – hypotheses, methods and empirical results. English version of 1984a, in *Theory and decision*, 1987.
Allais, M. 1984c. Determination of cardinal utility according to an intrinsic invariant model. Abridged version of 1984a, in *Recent developments in the foundations of utility and risk theory*, ed. L. Daboni et al., 83–120. Dordrecht: Reidel, 1985.
Allais, M. 1985. Three theorems on the theory of cardinal utility and random choice. In *Essays in honour of Werner Leinfellner*, ed. H. Berghel. Dordrecht: Reidel, 1986.
Allais, M., and O. Hagen, eds. 1979. *Expected utility hypotheses and the Allais' Paradox; contemporary discussions and rational decisions under uncertainty with Allais' rejoinder*. Dordrecht: Reidel.
Amihud, Y. 1974. Critical examination of the new foundation of utility. *In Allais and Hagen* 1979: 149–160.
Amihud, Y. 1977. A reply to Allais. In Allais and Hagen (1979), 185–190.
Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. Trans. as 'Exposition of a new theory on the measurement of risk'. *Econometrica* 22 (1954): 23–36.
de Finetti, B. 1977. A short confirmation of my standpoint. In Allais and Hagen (1979), 161.
Friedman, M., and J.L. Savage. 1948. The utility analysis of choices involving risk. *Journal of Political Economy* 56 (August): 279–304.
MacCrimmon, K., and S. Larsson. 1975. Utility theory: Axioms versus paradoxes. In Allais and Hagen (1979), 333–409.
Marschak, J. 1950. Rational behavior, uncertain prospects and measurable utility. *Econometrica* 18 (2): 111–141.
Marschak, J. 1951. Why 'should' statisticians and businessmen maximize moral expectation? In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. Berkeley: University of California Press.
Marschak, J. 1977. Psychological values, and decision makers. In Allais and Hagen (1979), 163–175.
Morgenstern, O. 1976. Some reflections on utility. In Allais and Hagen (1979), 175–183.
Samuelson, P. 1952. *Utility, preference and probability*. International Conference on Risk, Centre National de la Recherche Scientifique, Paris, May 1952. *Colloques Internationaux XL, Econométrie*, Paris (1953), 141–150.
Savage, L. 1952. *An axiomatization of reasonable behavior in the face of uncertainty*. International Conference on Risk, Paris, May 1952. Centre National de la Recherche Scientifique, *Colloques Internationaux XL, Econométrie*, Paris (1953), 29–33.
Savage, L. 1954. *The foundations of statistics*. New York: Wiley.
von Neumann, J., and O. Morgenstern. 1947. *Theory of games and economic behavior*. 2nd ed. Princeton: Princeton University Press.

Allais, Maurice (Born 1911)

Bernard Belloc and Michel Moreaux

Keywords

Allais, M.; Arbitrage; Bernoullian principle; Boiteux, M.; Business cycles; Capitalistic optimum; Choice under risk; Debreu, G.; Desrousseaux, J.; Divisia, F. J. M.; Dupuit, A.-J.-E.; Expected utility hypothesis; Friedman, M.; Functional analysis; Golden rule of accumulation; Hicks, J. R.; Hyperinflation; Industrial policy; Interdependence; Intertemporal general equilibrium; Intertemporal optimality; Jevons, W.S.; Malinvaud, E.; Marschak, J.; Morgenstern, O.; Neoclassical growth theory; Optimum population; Probability; Samuelson, P.A.; Savage, L. J.; Steady state equilibrium; Surplus; Tâtonnement; Uncertainty; Von Neumann, J.; Walras, L

JEL Classifications

B31

Maurice Allais was born on 31 May 1911, in Paris. Originally a student at the Ecole Polytechnique he moved later to the Ecole Nationale des Mines (ENMP hereafter). He gained the doctorate of engineering of the University of Paris in 1949. He is currently director of Research at the Centre National de la Recherche Scientifique (CNRS) and Professor of Economic Analysis at the ENMP. The CNRS awarded him a gold medal

in 1978, the first time this award was given to an economist. He was awarded the Nobel Prize for Economics in 1988.

His initial professional activity led him toward problems of applied economics and regulation. In France, the corps of mining engineers, one of the greatest branches of the civil service, is entrusted with the regulation of mining and energy and is very influential in the definition and control of public industrial policy. In some sense Allais's theoretical works are an attempt to find rational public economic public decisions. The title of his first book, *A la recherche d'une discipline économique. Première Partie: l'économie pure* (1943) is very significant in this respect. One feels in Allais's thought a deep reluctance to accept any theory which cannot be made operative (1978a). Thus a very important part of his activity, which will not be surveyed here, is devoted to applied economic studies, always, directly supported by a theoretical analysis (see 1954; 1956a; 1977). In the brilliant tradition of Dupuit, Colson and Divisia this aspect of Allais's work has been essential for the development of the school of French economist engineers. Allais educated several generations of researchers and public managers: M. Boiteux, G. Debreu and E. Malinvaud were among his students.

In the line of descent from Walras, Fisher and Pareto, Allais's theoretical contributions are basic in four fields: general equilibrium and optimal allocation of resources ('*rendement social'* or '*efficacité maximale'* in Allais's terminology), capital and growth, money and business cycle, risky choices.

Allais is primarily a theorist of interdependence and optimum. It is impressive to observe that the research programme defined at the start in Allais (1943) has been almost wholly fulfilled, even though some of the initial basic assumptions have been drastically revised. When published in 1943, Allais's book was one of the most complete reports on general equilibrium and optimum theories, comparable to Hicks's *Value and Capital* and Samuelson's *Foundations of Economic Analysis.* Let us emphasize its differences. Allais gives the earliest formalization of an intertemporal general equilibrium and, in particular, all the arbitrage conditions between capital goods and land are made explicit. Then, the first results on global stability of Walrasian *tâtonnement* are proved by means of Lyapunov's second method under assumptions equivalent to gross substitutability (see Negishi, *Econometrica* (1962), for a report in English). The book also contains a complete account of optimum theory in terms of distributable surpluses and a precise and correct statement of the two welfare theorems. Finally, Allais outlined a theory of optimum population. Later, Allais's opinion on the relevance of the Walrasian model changed markedly (1967b; 1968; 1971; 1981). He would now define a state of general equilibrium as a position in which no distributable surplus can be obtained, and describes the whole motion of the system as governed by the search for such surpluses. In some way this new view is a true merging of general equilibrium and optimum theories (1981).

His main contributions to capital and growth theory are expressed in Allais (1947; 1960; 1962). First, and sometimes with a lead of 15 years, he found most of the results of so-called neoclassical theory of growth, including the famous golden rule of accumulation. Allais worked out a complete theory of capitalistic processes with a rigorous formalization of the concept of characteristic function first proposed by Jevons in 1871, by which is meant the sequence of past expenditures on primary inputs which have generated the present national income. The systematic use of this concept allowed Allais to build up a theory of economic growth. But its use has been even more fruitful in the analysis of capitalistic efficiency. Allais proved in 1947 that, in a stationary state, a zero rate of interest maximizes real income. This is the first version of the golden rule of accumulation obtained by Phelps some 14 years later. In 1962 Allais widened this result and demonstrated that in steady states a capitalistic optimum is attained when the rate of interest is equal to the rate of growth (it is to be noted that Allais himself acknowledges that J. Desrousseaux had been the first to get this result in 1959, in a non-published paper). Thus Allais was completing his theory of optimal allocation of resources with a theory of capitalistic optimum.

To analyse intertemporal optimality, he assumes that each agent has preferences, on present and future consumption, possibly different in different periods. Hence it becomes possible to consider the psychological evolution of an individual over his lifetime, unlike the usual approach. In other respects Allais has been very careful to test the explicative power of his capitalistic optimum theory, by comparing the growth processes in different countries and trying to evaluate in every case the gap between the capitalistic optimum and the real state of accumulation.

Allais must be also considered as a major actor in the revival of the quantity theory of money (1956b, c; 1965a; 1966; 1969; 1970; 1972; 1974). The reduced form of the model explaining the dynamics of national monetary expenditure is very similar to Cagan's contemporary formulation. But Allais claims that his model has very different foundations because it is supported by an alleged psychological law of the perception of time. The solutions of the integro-differential equation describing the evolution of income are shown to have three limit cycles, depending upon initial conditions. It is then possible to explain local stability of a steady state equilibrium, business cycles and hyperinflation state with the same basic model.

The last aspect of Allais's work concerns choice under risk (1953b, c; 1979). As usual, Allais's approach is both theoretical and empirical. He builds up his analysis on the basis of experimental psychological tests conducted in 1952 (see Allais 1953c, for a partial statement). For Allais the theory of choice under risk went, historically, through four steps. At first it was assumed that the mathematical expectation of the monetary gain was the natural evaluation of a lottery. Then the mathematical expectation of the gain in utility was used. The third step then considered subjective probabilities. The American school (Friedman, Marschak, von Neumann, Morgenstern, Samuelson and Savage) takes into account only these three steps. So Allais claims that a fourth step must be reached: the value of a lottery is a functional depending upon the probability density parameterized by the gains. In effect the expected utility hypothesis implies a special such functional, so this last step seems very natural. Allais systematically criticizes the axioms on which the Bernoullian principle is based. According to him such axioms cannot help to define rationality in an uncertain environment. Through convincing examples he specially refutes Savage's independence and Samuelson's substitutability axioms. The major argument is in short that in the neighbourhood of certainty, a rational agent will prefer absolute safety. Then Allais proposes an alternative definition of rationality in risky situations: the set of choices must be ordered, an absolute preference axiom must be satisfied (that is, if a lottery gives in every case larger gains than another, then any agent will prefer the first one) and only objective probabilities must be considered. The first two axioms seem quite reasonable and it is difficult, according to Allais, to disprove the last one. But it is clear that a decision rule following the Bernoullian principle cannot be deduced from these three axioms. They imply the use of a functional of more general form than the mathematical expectation of the psychological evaluation of gains. In fact Allais argues that the Bernoullian principle only takes into account the dispersion of the gains whereas the dispersion of their psychological values is pertinent.

Finally, Allais applies his theory of behaviour under uncertainty to a general equilibrium model (1953a). He demonstrates this through an example where a competitive allocation of risks leads to an optimal allocation of resources, and where such an allocation can be obtained as a competitive equilibrium with an appropriate redistribution of initial endowments.

See Also

► Expected Utility Hypothesis

Selected Works

For a complete record of Allais's work on the period 1943–78 and an analysis by Allais himself, see Allais (1978a, b, c).

1943. *A la recherche d'une discipline économique. Première partie: I'économie pure*. Paris: Ateliers Industria.

1947. *Economie et intérêt.* Paris: Imprimerie Nationale.

1953a. L'extension des théories de l'équilibre économique général et du rendement social au cas du risque. *Econometrica* 21: 269–290.

1953b. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'Ecole Américaine. *Econometrica* 21: 503–546. Trans. in *Expected utility hypotheses and the allais paradox,* ed. M. Allais and O. Hagen. Dordrecht: Reidel, 1979.

1953c. La psychologie de l'homme rationnel devant le risque. La théorie et l'expérience. *Journal de la Société de Statistique de Paris* January: 47–72.

1954. Méthode d'évaluation des perspectives économiques de la recherche minière sur des grands espaces. Application au Sahara Algérien. Paris, mimeo. Trans. in *Management Science* 3(4) (1957), 285–347.

1956a. *La gestion des houillères nationalisées et la théorie économique*. Paris: Imprimerie Nationale.

1956b. Explication des cycles économiques par un modèle non linéaire à régulation retardée. *Metroeconomica* 8: 4–83.

1956c. Explication des cycles économiques par un modèle non linéaire à retardée. Mémoire complémentaire. In *Les modèles dynamiques en économétrie,* Collection des Colloques Internationaux du CNRS, vol. 52. Paris: CNRS.

1960. L'influence du coefficient capitalistique sur le revenu réel par tête. *International Statistical Institute Bulletin* 38(2): 3–27.

1962. The influence of the capital-output ratio on real national income. *Econometrica* 30: 700–728.

1965a. Reformulation de la theórie quantitative de la monnaie. *Bulletin SEDEIS* 928(Supplement).

1965b. The role of capital in economic development. In *The economic approach to economic development*. Amsterdam: North-Holland.

1966. A restatement of the quantity theory of money. *American Economic Review* 56: 1123–1157.

1967a. Some analytical and practical aspects of the theory of capital. In *Activity analysis in the theory of growth and planning*, ed. E. Malinvaud and M. Bacharach. London: Macmillan.

1967b. *Les fondements du calcul économique,* vol. 1. Paris: Ecole Nationale Supérieure des Mines (mimeo).

1968. *Les fondements du calcul économique,* vols. 2 and 3. Paris: Ecole Nationale Supérieure des Mines (mimeo).

1969. Growth and inflation. *Journal of Money, Credit and Banking* 1: 355–426.

1970. A reply to Michael R. Darby's comment on Allais's restatement of the quantity theory. *American Economic Review* 60: 447–456.

1971. Les théories de l'équilibre économique général et de l'efficacité économique. Imapasses récentes et nouvelles perspectives. *Revue d'Economie Politique,* May. Trans. in *Equilibrium and didsequilibrium in economic theory,* ed. G. Schwödiauer. Dordrecht: Reidel.

1972. Forgetfulness and interest. *Journal of Money, Credit and Banking* 4(1, Part I): 40–73.

1974. The psychological rate of interest. *Journal of Money, Credit and Banking* 6: 285–331.

1975. The hereditary and relativisitc formulation of the demand for money. Circular reasoning or a real structural relation? (A reply to Scadding's Note.) *American Economic Review* 65: 454–464.

1977. *L'impôt sur le capital et la réforme monétaire*. Paris: Hermann.

1978a. *Contributions à la science économique. Vue d'ensemble 1943–1978*. Paris: Centre d'Analyse Economique.

1978b. *Principaux ouvrages et mémoires 1943–1978.* Paris: Centre d'Analyse Economique.

1978c. *Titres et travaux scientifiques.* Paris: Centre d'Analyse Economique.

1979. The so-called allais paradox and rational decisions under uncertainty. In *Expected utility hypotheses and the allais paradox*, ed. M. Allais and T. Hagen. Dordrecht: Reidel.

1981. La théorie générale des surplus. vol. II, Modèle économie et sociétés. *Cahiers de l'ISMEA,* Series EM no. 9.

Allen, George Cyril (1900–1982)

Audrey Donnithorne

George Allen was born on 28 June 1900 at Kenilworth, Warwickshire, England and died at Oxford on 31 July 1982. His upbringing and education in Coventry and then in Birmingham, influenced the choice of topic – the industrial development of the Black Country – for his first book on what was to become one of his main professional interests, the history and organization of British industry. His second major interest, the economy of Japan, came from the 3 years he spent as a young man teaching at Nagoya, Japan. He returned to take up a post at Birmingham University from which, while still in his twenties, he was appointed to a chair at Hull and later, at Liverpool. In World War II he worked at the Board of Trade and the Ministry of Economic Warfare. His wartime activities included playing a key part in the concentration of civilian industry and, with Hugh Gaitskell, writing the first paper on postwar policy on monopolies and restrictive practices. Later, he spent 6 months at the Foreign Office to advise on the economic reconstruction of Japan under the allied occupation. From 1947 to 1967 Allen headed the Department of Political Economy at University College, London, while continuing his participation in practical affairs as a member of the Monopolies (and Restrictive Practices) Commission and of other official bodies.

For Allen, economics was part of the study of man, not the application of specialized techniques. Therefore he favoured an historical approach, emphasizing the importance of institutional and social factors. At a time when government control and economic planning were widely considered as panaceas, Allen remained sceptical. Of his 17 books, perhaps the best known were *British Industries and their Organisation* (1933, revised edition 1970), *A Short Economic History of Japan* (1946, revised edition 1981) and *Monopoly and Restrictive Practices* (1968).

Selected Works

1933. *British industries and their organisation*. 5th edn, London: Longmans, 1970.
1946. *A short economic history of modern Japan*. 4th edn, London: Macmillan, 1981.
1968. *Monopoly and restrictive practices*. London: Allen & Unwin.

Allen, Roy George Douglas (1906–1983)

J. R. N. Stone

Keywords

Allen, R. D. G.; Bowley, A. L.; Econometric society; Econometrics; Hicks, J. R.; Index numbers; Mathematics and economics; National accounting; Positive economics; Royal statistical society; Statistics and economics; Utility; Value

JEL Classifications

B31

Allen was born on 3 June 1906 at Stoke-on-Trent, and died on 29 September 1983 at Southwold. He was knighted in 1966 and made a Fellow of the British Academy in 1952. He was educated at the Royal Grammar School, Worcester, and Sidney Sussex College, Cambridge. From 1928 he was assistant, then lecturer, then reader in economic statistics at the London School of Economics, becoming professor of statistics in 1944 and emeritus professor in 1973.

During the war, he was a statistician in H.M. Treasury from 1939 to 1941; from 1941 to 1942 he was Director of Records and Statistics for the British Supply Council in Washington, and from 1942 to 1945, he became British Director of Research and Statistics for the Combined Production and Resources Board in Washington. His other principal activities were as statistical adviser

for H.M. Treasury (1947–1948); member of the Air Transport Licensing Board (1960–1972); and member of the Civil Aviation Authority (1972–1973). He was President of the Econometric Society in 1951 and President of the Royal Statistical Society in 1969–1970. He was also consultant to many international and professional organizations.

Allen was an economic statistician, mathematical economist and econometrician of exceptional competence and breadth of knowledge. His early and most original research, carried out in part with J.R. Hicks and A.L. Bowley, was on the theory of value, utility and consumers' behaviour: for example, Hicks and Allen (1934), Allen (1935), and Allen and Bowley (1935), the last an outstanding work on the econometrics of family budgets.

In the late 1930s he embarked on a series of successful textbooks based on his lectures. His *Mathematical Analysis for Economists* (1938) was intended to help students of economics whose training in mathematics was typically much less thorough than it is now. After the war, in addition to numerous papers on economic and statistical topics, including one reflecting his wartime work in Washington (Allen 1946), and a compilation of papers on international trade statistics (Allen and Ely 1953), he continued the good work begun in 1938 with a succession of books on macroeconomics and the mathematical and statistical tools required in its study. Thus *Statistics for Economists* (1949) is an introduction to statistical methods in their application to economic material; *Mathematical Economics* (1956) is a text on economic theory, written in mathematical terms, which takes account of the growth of econometrics and the use of increasingly sophisticated mathematics by economists; *Basic Mathematics* (1962) provides a general introduction to mathematical ideas, applicable in both the natural and the social sciences; *Macro-Economic Theory* (1967) treats deterministic models from a positive rather than an optimizing or policy-oriented point of view; his 1975 work deals comprehensively with the design, construction and use of index numbers, paying full attention to both the economic and the statistical aspects of the subject; his last book (1980) is an introduction to national accounting, concentrating on the main aggregates at current and constant prices and illustrated by means of recent British official estimates.

Allen was an assiduous disseminator of ideas. His textbooks were translated into many languages and he continued to lecture until shortly before his death. As head of the Statistics Department of the LSE he was instrumental, with the help of M.G. Kendall, in expanding it from a staff of five in 1944 to one of 28, of whom seven were professors.

Selected Works

1934. (With J.R. Hicks.) A reconsideration of the theory of value: parts I and II. *Economica,* February, 52–76; May, 196–219.

1935. A note on the determinateness of the utility function. *Review of Economic Studies* 2, February, 155–8.

1935. (With A.L. Bowley.) *Family Expenditure.* London: P.S. King.

1938. Mathematical Analysis for Economists. London: Macmillan.

1946. Mutual aid between the US and the British Empire, 1941–45. *Journal of the Royal Statistical Society* 109, 243–71.

1949. *Statistics for Economists.* London: Hutchinson.

1953. (With J.E. Ely, eds.) *International Trade Statistics.* London: Chapman and Hall.

1956. *Mathematical Economics.* London: Macmillan.

1962. *Basic Mathematics.* London: Macmillan.

1967. *Macro-Economic Theory.* London: Macmillan.

1975. Index Numbers in Theory and Practice. London: Macmillan.

1980. An Introduction to National Accounts Statistics. London: Macmillan.

Bibliography

Cairncross, A. 1985. Roy Allen, 1906–1983. *Proceedings of the British Academy* 70: 379–385.

Grebenik, E. 1984. Roy George Douglas Allen, 1906–1983. *Journal of the Royal Statistical Society* 147: 706–707.

Almon Lag

Roger N. Waud

The Almon distributed lag, due to Shirley Almon (1965), is a technique for estimating the weights of a distributed lag by means of a polynomial specification.

Consider the distributed lag model,

$$y_t = w_0 x_t + \ldots + w_n x_{t-n} + \varepsilon_t \qquad (1)$$

where y_i is the value of the dependent variable at time $t; x_t, x_{t-1}, \ldots, x_{t-n}$ are the values of the regressor x at times, $t, t-1, \ldots t-n$; and ε_t is the value of the disturbance ε at time t. The dependent variable y is influenced by the regressor x both contemporaneously and with a lag of up to n time periods. If the lag length, n, is finite and less than the number of observations, the regression coefficients w_i can be estimated by ordinary least squares (OLS).

It is often the case, however, that there is a high degree of multicollinearity among the regressors $x_t, \ldots, x_{t-n}$ so that most or all of the estimated regression coefficients are statistically insignificant, and powerful inferences about the true weights are impossible. This problem can be circumvented by introducing *a priori* information into the estimation procedure, typically by imposing restrictions on the true weights. If the restrictions are valid, the estimates of the weights will be unbiased, consistent, *and* more efficient than the OLS estimates. Similarly, the tests of hypotheses about the true weights will be valid *and* more powerful than the tests based on OLS estimation.

The Almon lag technique introduces *a priori* information by estimating the distributed lag model (1) subject to the restriction that the weights lie on a polynomial of degree p,

$$w_i = \lambda_0 + \lambda_1 i + \lambda_2 i^2 + \cdots + \lambda_p i^p, \qquad (2)$$

$i = 0, 1, \ldots, n; p \leq n$. This reduces the number of parameters from $n+1 (w_0, w_1, \ldots, w_n)$ to $p+1$ $(\lambda_0, \lambda_1, \ldots, \lambda_p)$. (A very readable description of the procedure for estimating the 'new' parameters (λ_0, λ_1, …, λ_p) and transforming these into estimates of the original weights (w_1, w_2, …, w_n) is provided by Kmenta (1971, pp. 492–3).) As with any *a priori* restriction, the restriction that the weights lie on a polynomial will lead to more efficient and more powerful tests if the restriction is valid, but will give biased and inconsistent and invalid tests if the restriction striction is false. Following are some important caveats to be borne in mind when using the Almon technique.

The Presence or Absence of a Lag Is Not a Testable Proposition When the Almon Lag Technique Is Used

Suppose no lag is present so that x affects y only instantaneously; $w_0 \neq 0$ but $w_1 = w_2 = \cdots = w_n = 0$.. Since a polynomial of degree p can equal zero in only p places (unless it is identically zero), any choice of $p < n$ involves a specification error; the n zeros w_1, w_2, …, w_n cannot lie on a polynomial of degree $p < n$. Therefore if the Almon lag technique is used in this case, the results will suggest the presence of a lag even though there is none.

The Use of End-Point Constraints

It has been a rather common practice among users of the Almon technique to impose one or both end-point constraints

$$w_{-1} = 0; \quad w_{n+1} = 0 \qquad (3)$$

in estimation. In terms of (2) this involves the following restrictions on the λs:

$$\lambda_0 - \lambda_1 + \lambda_2 - \cdots \pm \lambda_p = 0 \qquad (4)$$

$$\lambda_0 + (n+1)\lambda_1 + (n+1)^2 \lambda_2 + \cdots + (n+1)^p \lambda_p = 0. \qquad (5)$$

The imposition of (4) and (5) increases the efficiency of estimation if the restrictions are

true, but gives biased and inconsistent estimates if they are false. In general, however, there are no convincing reasons for imposing these constraints. For example, it is tempting to argue that $w_{-1} = 0$ because it is the coefficient on x_{t+1} and x_{t+1} does not affect y_t. By the same logic one would conclude $0 = w_{-2} = w_{-3} = w_{-4} = \cdots$. However, this is not possible. If the weights w_i do in fact lie on a polynomial of degree p, no more than p of them can equal zero. This illustrates why one should be concerned only with the weights $w_0, w_1, \ldots, w_n$ – the behaviour of the polynomial outside this range is irrelevant.

Choosing the Lag Length and Polynomial Degree

Understating the length (choosing n less than the true lag length) is a specification error which results in biased and inconsistent estimates and invalid tests. A specification error is also committed by overstating the lag length. This occurs whenever the lag length is overstated by more than p minus the number of endpoint constraints because a p-degree polynomial can have only p zeros. Choosing a small value of p increases the possible efficiency gain from use of the Almon technique, but also makes specification error more likely. However, if p is not considerably less than n, using the technique may be pointless since the results will strongly resemble the OLS results; when $p = n$ the estimates are the same as OLS. A discussion of the procedures for testing for appropriate lag length and degree of polynomial, along with relevant literature citations, can be found in Judge et al. (1980, pp. 645–51).

See Also

► Multivariate Time Series Models

References

Almon, S. 1965. The distributed lag between capital appropriations and expenditures. *Econometrica* 33(1): 178–196.

Judge, G., W. Griffiths, R. Hill, and T. Lee. 1980. *The theory and practice of econometrics*. New York: Wiley.

Kmenta, J. 1971. *Elements of econometrics*. New York: Macmillan.

Almon, Shirley Montag (1935–1975)

Roger N. Waud

Shirley Almon was born on 6 February 1935, in Saxonburg, Pennsylvania and died on 27 September 1975, in College Park, Maryland. She graduated from Goucher College in Baltimore, Maryland, in 1956, and received her Ph.D. from Harvard University in 1964. The essence of her Ph.D. dissertation was published in *Econometrica* (1965), a frequently cited article that introduced a new statistical technique for estimating distributed lags. This technique, now commonly known as the Almon lag, has been widely used in numerous econometric studies.

Almon worked at various times as an economist at the Women's Bureau in Washington, DC, at the National Bureau of Economic Research, at the Federal Reserve Bank of San Francisco, and at the Federal Reserve Board in Washington, DC. She taught elementary economics, industrial organization, and statistics at Wellesley College and Harvard University before joining the staff of the President's Council of Economic Advisers in 1966, where she continued until the onset of a brain tumour, discovered in 1967, ended her brief but significantly productive career.

Selected Works

1965. The distributed lag between capital appropriations and expenditures. *Econometrica* 33:178–196.

1968. Lags between investment decisions and their causes. *Review of Economics and Statistics* 50:193–206.

A

Altruism

Peter J. Hammond

The French term 'altruisme' was introduced by Auguste Comte (1830–42) to signify devotion to the welfare of others, especially as a principle of action. It is closely related to concepts such as benevolence and unselfishness. It has long attracted the interest of moral philosophers (see e.g. Nagel 1970; Milo 1973; Roberts 1973; Collard 1978; Margolis 1982). Rescher (1975, p. 11) categorizes it as one of the 'modalities' of unselfishness. Numerous social scientists in many fields, including sociobiology, have been interested in altruistic behaviour as helping to assure species and gene survival (Becker 1976; Collard 1978, ch. 5). While some economists have participated in such research, more have naturally concentrated upon the implications of altruism for economic outcomes–in particular, the allocation of resources and the distribution of income.

Altruistic Preferences and Utilities

Most of the problems presented by altruism are adequately captured in a simple model with n individuals who each consume a single transferable good – perhaps a Hicks composite commodity, because relative prices are fixed. So an economic allocation is described by an income distribution vector $\boldsymbol{y}$ in $\mathbb{R}^n_+$ whose typical non-negative component y_i denotes the income of person i. Even intergenerational altruism can be discussed in such a framework, with y_i denoting wealth, provided that capital markets are perfect and no transfers occur which affect real interest rates, and provided that we ignore the special problems that arise when both the time horizon and the number of individuals are infinite.

If individual i has selfish preferences, then income distribution $\mathbf{y}$ is preferred to $\mathbf{y}'$ if and only if $y_i > y'_i$ so that i has more income. But altruistic preferences allow $\mathbf{y}$ to be preferred to $\mathbf{y}'$ even if $y_i < y'_i$ provided enough other individuals j have gains $y_j - y'_i$ which are large enough to overcompensate. Thus altruistic preferences can be quite a general (complete and transitive) ordering $\succsim i$ on $\mathbb{R}^n_+$.

Some more care is needed here, however. Economists usually identify 'welfare' with 'preferences' and assume that it can be represented by a welfare function $w_i(\boldsymbol{y})$ on $\mathbb{R}^n_+$ which increases as $\boldsymbol{y}$ becomes more preferred. Recalling that altruism is regard for others' *welfare* then suggests that i must want to maximize a function of the form $w_i = \phi_i(y_i, \boldsymbol{w}_{-i})$ where $\mathbf{w}_{-i}$ denotes the vector of welfare levels w_j $(j \neq i)$ with i excluded, and where ϕ_i is increasing in every other w_j. Given the income distribution $\mathbf{y}$, finding the individual welfare levels $w_i(\boldsymbol{y})$ requires solving the n simultaneous equations:

$$w_i = \phi(y_i, \boldsymbol{w}_{-i}) \tag{1}$$

for every i. So each $w_i(y)$ is only well-defined provided that these equations have a unique solution. Becker (1974, pp. 1076–7) amongst others discusses this problem – for a special Cobb–Douglas case with two individuals. Assume that Eq. 1 does have a unique solution for every $\boldsymbol{y}$ in $\mathbb{R}^n_+$, though this is by no means innocuous. In particular, taking the total differential of Eq. 1 gives:

$$dw_i - \sum_{j \neq i} \phi_{ij} dw_j = \phi_{ii} dy_i \tag{2}$$

and the matrix formed by the coefficients of each dw_j on the left-hand side of each equation in (Eq. 2) must be invertible (see Kolm 1969, pp. 153–4).

Pareto Inefficient Redistribution

When everybody's altruistic utility function $w_i(y)$ depends upon the incomes of all, it seems obvious that there are externalities likely to cause Pareto

inefficiency. Unlike standard externalities, however, individuals can translate their altruism into action by giving income away to anyone they want to. Let t_{ij} (≥ 0) denote the transfer made by i to j. Then, assuming that each w_j is differentiable, and recognizing the non-negativity constraints that prevent people taking income from others, transfers occur until the following first order conditions are satisfied for every i, j with $i \neq j$:

$$w_{ij} \leq w_{ii}, t_{ij} \geq 0 \text{ and } t_{ij}\left(w_{ii} - w_{ij}\right) = 0 \qquad (3)$$

where w_{ii} denotes $\partial w_i/\partial y_j$. Thus $w_{ii} = w_{ij}$ unless the constraint $t_{ij} \geq 0$ binds, when one can have $w_{ii} < w_{ij}$, with i valuing his own income more than j's at the margin.

First order conditions for Pareto efficiency, on the other hand, require the existence of marginal welfare weights $\beta_i i = (1\ to\ n)$ such that, for every pair of individuals j, k:

$$\sum_i \beta_i \left(w_{ij} - w_{ik}\right) = 0 \qquad (4)$$

so that the marginal social benefit of \$1 for j is equal to that of \$1 for k. This presumes an interior distribution in which all have income.

Now suppose that Eq. 3 is satisfied at a distribution $\mathbf{y}^*$ in which no individual wants to take income from anybody else. Then $w_{ii} = w_{ij}$ for all i, j and the efficiency conditions (Eq. 4) are satisfied! But Winter (1969) notices how alleviating poverty can create externalities of the kind that occur when public goods have to be provided by private individuals. After all, a poor person is likely to benefit by receiving income from the rich, even if he is altruistic to the rich. Then $w_{ij} < w_{ii}$, where i is the poor person and j the rich. So Eq. 4 may well be violated. This is especially clear in Arrow (1981), where every individual's altruistic utility takes the form:

$$w_i(\boldsymbol{y}) \equiv u_i(y_i) + \sum_{j \neq 1} \upsilon\left(y_j\right) \qquad (5)$$

and $y_i = y_j$ implies $u_i'(y_i) > \upsilon'\left(y_j\right)$. Arrow shows that, excluding trivial equilibria in which no voluntary redistribution at all takes place, redistribution is only Pareto efficient in a very special case when there is just one rich giver – e.g. the two person case by Hochman and Rodgers (1969). Obviously giving is then not a public good. But as soon as there are two or more givers, Pareto inefficiency is inevitable in Arrow's model at least (see also Bergstrom 1970; Nakayama 1980).

Policy Relevance

If altruistic preferences make transfers to the poor a public good, this is a prima facie argument for public intervention to redistribute income. Yet this argument has been contested. It has even been claimed that redistributive policy is powerless because it merely substitutes for private charity. In Barro (1974), the issue is obscured by dynamic considerations and the fact that 'charity' takes the form of bequests to one's heirs. Public debt becomes irrelevant because its effects are totally offset by bequests. This presumes, however, that nobody wishes to make negative bequests, because otherwise the national debt is a way of reproducing the effects of negative bequests. Were Barro's arguments correct, Bernheim and Bagwell (1985) show that then many other policy instruments would also become ineffective; even distortionary income taxes could be offset by reducing bequests in order to pay them. Just as Barro's neutrality proposition fails when agents could gain from making negative bequests if they were allowed, so redistributive policies are effective *precisely* when the poor can gain by further transfers from the rich which the rich are unwilling to make because of the public good problem. This is true even when the poor feel altruistic toward the rich.

There are some special cases where neutrality does hold and policy is irrelevant. One is with just one giver, which is the Arrow (1981) sufficient condition for efficiency. Another is Becker's (1974, p. 1080, 1981) household with a head who is wealthy enough to want to control the intrafamily distribution of income. Then the 'rotten kid' theorem has the activities of selfish children completely offset by transfers from the head,

provided that the household head is able to retain control even if he should die first (Hirshleifer 1977). Becker never considers, however, a family with two heads, for which the rotten kid theorem would fail in general, with each head providing too little support for efficiency. Warr (1982, 1983) also argues that policy is irrelevant, but analyses only first order conditions like Eq. 3 without any inequalities, and so fails to consider the likely case in which charity is insufficient to make the poor want no more transfers from the rich. Bergstrom and Varian (1985) and Bergstrom et al. (1986) provide further discussion.

Is Charity Public Good?

The flawed policy irrelevance argument is not the only way to contest treating redistribution as a public good when there is altruism. A better argument is due to Sugden (1982, 1983, 1985) who questions whether individuals' altruistic behaviour maximizes a utility function which can be expressed solely as a function of the income distribution. Suppose that A likes to give 10% of his marginal income to charity C. Suppose too that B gives $10 less to charity C than before. Then this is like a $10 fall in A's total income, leading to a $1 drop in the amount A wants C to receive in total, so A increases his giving to C by $9 in order to bring this about. Conversely, if B gives $10 more to charity C, then A will reduce his giving by $9. This is a general feature of privately provided public goods: the more one person gives, the less others will want to. In the case of charity, however, such negative covariation between different people's giving seems implausible. It would imply (Sugden 1983) that the main beneficiaries of a new gift to a charity are those other givers who respond by reducing their gifts to that charity! Sugden concludes that givers value charity per se as well as for the help it gives the recipients-a possibility discussed earlier by Arrow (1974), amongst others. Then each person's giving becomes a separate *private* good, and the public good argument for replacing charity by tax-financed transfer programmes becomes less convincing.

Charity: Real or Apparent Altruism?

An obvious explanation of behaviour such as charitable giving is altruistic preferences. Yet it is not the only possible explanation. As just discussed, gifts may be made for their own sake as well as because of an altruistic regard for the recipient's welfare. They may also reflect egoistic cooperative behaviour, however, as discussed by Boulding (1973), Arrow (1974), and Hammond (1975), amongst economists, and by many sociobiologists – as has been pointed out by Becker (1976), Kurz (1977, 1978), etc. If persons A and B are in continual contact, both may gain from reciprocal cooperation as a form of mutual insurance. And if there is an infinite chain A_1, A_2, A_3,... with person A_n in contact with A_{n+1}, all can gain from maintaining cooperation into the indefinite future. Genes which promote such cooperation enhance their prospects of long-run survival. Maintaining such cooperation requires deviants to be punished suitably. But apparently altruistic behaviour emerges from entirely selfish preferences. The same is true when it is clear to all members of a group that *one* of their members must act for their mutual benefit, although there may be a costly or dangerous delay before the apparently altruistic behaviour emerges, as in Bliss and Nalebuff's (1984) model of brinkmanship. Finally, Sugden's (1984) theory of reciprocity is an interesting recent explanation of apparently altruistic behaviour which is really selfish at bottom.

Is Altruism Relevant?

It has just been seen that altruistic preferences may be unnecessary to explain apparently altruistic behaviour. This limits the relevance of altruism for positive economics, though one cannot deny that some behaviour is indeed motivated by altruism in the sense of devotion to the welfare of others.

A more controversial claim is that altruism also has limited relevance for normative economics. This issue is addressed by Barry (1965, p. 65) because altruistic regard for others is an instance of a 'publicly oriented want', which 'carries a

claim to satisfaction only as being a want for what ought to be done anyway', thus people's altruistic preferences are irrelevant in determining what should be the distribution of income, except in so far as they correspond to what is anyway ethically appropriate. In the language of welfare economics, each individual's welfare corresponds only to that person's 'privately-oriented' or selfish preferences. Altruism is therefore excluded, because it is regard for *others*' welfare. That is not to deny that welfare-relevant externalities may arise if, for instance, a rich person experiences revulsion on being confronted with extreme poverty. But avoiding such revulsion is *not* altruism so much as selfish behaviour.

A related reason for excluding altruistic preferences from welfare is to avoid undesirable double counting. Suppose that A is an altruist with utility $u(y^A) + (1/2)u(y^E)$ for the distribution of income between A and E, an egoist. Suppose that E, however, has a selfish utility function $u(y^E)$. Adding utilities then gives $u(y^A) + (3/2)u(y^E)$, with greater weight for the marginal utility of the undeserving egoist's income than for the deserving altruist's. A more appropriate welfare function is $u(y^A) + u(y^E)$ which disregards A's altruism and just adds selfish utilities. The main role of altruism in welfare economics is to help determine ethical views, not to determine individual welfare. Concepts of Pareto efficiency which include altruism in individual welfare have little normative significance, as do the alleged 'Pareto efficient' income redistributions. Altruistic behaviour often helps to promote social welfare, but it may not if the altruism happens to be directed toward those whose income should receive only small weight in the social welfare function.

See Also

- ▶ Envy
- ▶ Equity
- ▶ Externalities
- ▶ Family
- ▶ Gifts
- ▶ Public Goods
- ▶ Wicksteed, Philip Henry (1844–1927)

References

Arrow, K.J. 1974. Gifts and exchanges. *Philosophy and Public Affairs* 1: 343–362. reprinted as pp. 13–28 of Phelps (1975).

Arrow, K.J. 1981. Optimal and voluntary income distribution. In *Economic welfare and the economics of soviet socialism: Essays in honour of Abram Bergson*, ed. S. Rosefield, 267–288. Cambridge: Cambridge University Press. reprinted as ch. 15 of Collected Papers of Kenneth J. Arrow, Vol. 1: Social choice and justice. Cambridge, MA: Belknap Press.

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82: 1095–1117.

Barry, B.M. 1965. *Political argument*. London: Routledge & Kegan Paul.

Becker, G.S. 1974. A theory of social interactions. *Journal of Political Economy* 82: 1063–1091.

Becker, G.S. 1976. Altruism, egoism and genetic fitness: Economics and sociobiology. *Journal of Economic Literature* 14: 817–826.

Becker, G.S. 1981. Altruism in the family and selfishness in the market place. *Economica* 48: 1–16.

Bergstrom, T.C. 1970. A 'Scandinavian consensus' solution for efficient income distribution among nonmalevolent consumers. *Journal of Economic Theory* 2: 383–398.

Bergstrom, T.C., and H.R. Varian. 1985. When are Nash equilibria independent of the distribution of agents' characteristics? *Review of Economic Studies* 52: 715–718.

Bergstrom, T.C., L. Blume, and H.R. Varian. 1986. On the private provision of public goods. *Journal of Public Economics* 29: 25–49.

Bernheim, B.D., and Bagwell, K. 1985. Is everything neutral? Mimeo, Stanford University; forthcoming in the *Journal of Political Economy.*

Bliss, C.J., and B. Nalebuff. 1984. Dragon-slaying and ballroom dancing: The private supply of a public good. *Journal of Public Economics* 25: 1–12.

Boulding, K.E. 1973. *The economy of love and fear: A preface to grants economics*. Belmont: Wadsworth.

Collard, D.A. 1978. *Altruism and economy: A study in non-selfish economics*. Oxford/New York: Martin Robertson/Oxford University Press.

Comte, A. 1830–42. *Cours de philosophie positive,* 6 vols. Paris: Bachelier.

Hammond, P.J. 1975. Charity: Altruism or cooperative egoism? In Phelps (1975), 115–131.

Hirshleifer, J. 1977. Shakespeare vs. Becker on altruism: The importance of having the last word. *Journal of Economic Literature* 15: 500–502.

Hochman, H.M., and J.D. Rodgers. 1969. Pareto optimal redistribution. *American Economic Review* 59: 542–557.

Kolm, S.-C. 1969. The optimal production of social justice. In *Public economics*, ed. J. Margolis and H. Guitton, 145–200. London: Macmillan.

Kurz, M. 1977. Altruistic equilibrium. In *Economic progress, private values, and public policy: Essays in*

honor of William Fellner, ed. B. Balassa and R. Nelson, 177–200. Amsterdam: North-Holland.
Kurz, M. 1978. Altruism as an outcome of social interaction. *American Economic Review: Papers and Proceedings* 68: 216–228.
Margolis, H. 1982. *Selfishness, altruism, and rationality: A theory of social choice*. Cambridge: Cambridge University Press.
Milo, R.D. (ed.). 1973. *Egoism and altruism*. Belmont: Wadsworth.
Nagel, T. 1970. *The possibility of altruism*. Oxford: Clarendon Press.
Nakayama, M. 1980. Nash equilibria and Pareto optimal income redistribution. *Econometrica* 48: 1257–1263.
Phelps, E.S. (ed.). 1975. *Altruism, morality and economic theory*. New York: Russell Sage Foundation.
Rescher, N. 1975. *Unselfishness: The role of the vicarious affects in moral philosophy and social theory.* Pittsburgh: University of Pittsburgh Press.
Roberts, T.A. 1973. *The concept of benevolence: Aspects of eighteenth-century moral philosophy.* London: Macmillan.
Sugden, R. 1982. On the economics of philanthropy. *Economic Journal* 92: 341–350.
Sugden, R. 1983. *Who cares? An economic and ethical analysis of private charity and the welfare state*. London: Institute of Economic Affairs.
Sugden, R. 1984. Reciprocity: The supply of public goods through voluntary contribution. *Economic Journal* 94: 772–787.
Sugden, R. 1985. Consistent conjectures and voluntary contributions to public goods: Why the conventional theory does not work. *Journal of Public Economics* 27: 117–124.
Warr, P. 1982. Pareto optimal redistribution and private charity. *Journal of Public Economics* 19: 131–138.
Warr, P. 1983. The private provision of a public good is independent of the distribution of income. *Economics Letters* 13: 207–211.
Winter, S.G. 1969. A simple remark on the second optimality theorem of welfare economics. *Journal of Economic Theory* 1: 99–103.

Altruism in Experiments

James Andreoni, William T. Harbaugh and Lise Vesterlund

Abstract

We call an act altruistic when it is a sacrifice that benefits others. We discuss how experiments have demonstrated that altruistic choices appear to follow the same regularity conditions as those assumed for private goods. In particular they vary rationally in response to changes in prices and circumstances. We show how experiments have distinguished between different economic models of how concern for others enters utility functions, and have explored the implications of those models for charitable giving, labour markets, and trust. We also discuss the experimental evidence for differences in altruism by gender, and work on altruism's cultural, developmental, and neural foundations.

Keywords

Altruism; Altruism in experiments; Centipede game; Cooperation; Crowding out; Dictator game; Efficiency wages; Experiments; Gift exchange; Moral hazard game; Prisoner's Dilemma; Public goods games; Rawls, J.; Reciprocity; Repeated games; Reputation; Revealed preference; Trust game; Ulterior motives; Ultimatum game; Utilitarianism; Warm-glow

JEL Classifications

C9

Unlike experiments on markets or mechanisms, experiments on altruism are about an individual motive or intention. This raises serious obstacles for research. How do we define an altruistic act, and how do we know altruism when we see it?

The philosopher Thomas Nagel provides this definition of altruism: 'By altruism I mean not abject self-sacrifice, but merely a willingness to act in the consideration of the interests of other persons, without the need of ulterior motives' (1970, p. 79). Notice that there are two parts to this definition. First, the act must be in the consideration of others. It may or may not imply sacrifice on one's own part, but it does require that the consequences for someone else affect one's own choice. The second aspect is that one does not need 'ulterior motives' rooted in selfishness to explain altruistic behaviours. Of course,

ulterior motives may exist alongside altruism, but they cannot be the only motives.

If this is our definition of altruism, then how do we know altruism when we see it? The answer, unfortunately, is necessarily a negative one – we only know when we do not see it. Altruism is part of the behaviour that you cannot capture with a specifically defined ulterior motive. Experimental investigation of altruism is thus focused around eliminating any possible ulterior motives rooted in selfishness. One of the central motives that potentially confounds altruism is the warm-glow of giving, that is, the utility one gets simply from the act of giving *without* any concern for the interests of others (Andreoni 1989, 1990). While it is possible that warm-glow exists apart from altruism, it seems most likely that the two are complements – the stronger your desire to act unselfishly, the greater the personal satisfaction from doing so. Indeed, the two may be inextricably linked. Having a personal identity as an altruist may necessarily precede altruistic acts, and maintaining that identity can only come from actually being generous.

In what follows we will highlight the main experimental evidence regarding choices made in the interests of others, and the systematic attempts in the literature to rule out ulterior motives for these choices. Since these serious and repeated attempts to rule out ulterior motives have not been totally successful, the experimental evidence, like Thomas Nagel, favours the possibility of altruism.

Laboratory Experiments with Evidence of Altruism

In describing the games below, we adopt the convention of using Nash equilibrium to refer to the prediction that holds if all subjects are rational money-maximizers.

Prisoner's Dilemma

There have been thousands of studies using Prisoner's Dilemma (PD) games in the psychology and political science literatures, all exploring the stubborn nature of cooperation (Kelley and Stanelski 1970). Roth and Murnigham (1978) explored PD games under paid incentives and with a number of different payoff conditions. Their study confirmed to economists that cooperation is robust.

Sceptics noted, however, that cooperation need not be caused by altruism. First, inexperience and initial confusion may cause subjects to cooperate. Second, subjects in a finitely repeated version of the game may cooperate if they each believe there is a chance someone actually is altruistic. Behaviourally this 'sequential equilibrium reputation hypothesis' (Kreps et al. 1982) does not actually require subjects to be altruistic, but only that they believe that they are sufficiently likely to encounter such a person.

Andreoni and Miller (1993) explore these two factors by asking subjects to play 20 separate ten-period repeated PD games. A control treatment had subjects constantly changing partners, thus unable to build reputations. They find significant evidence for reputations, but that these alone cannot explain the level of cooperation, especially at the end of the experiment. Rather, they estimate that about 20 per cent of subjects actually need to be altruistic to support the equilibrium findings. This finding is corroborated in other repeated games, such as Camerer and Weigelt's (1988) moral hazard game, McKelvey and Palfrey's (1992) centipede game, and in a two-period PD of Andreoni and Samuelson (2006).

Public Goods

Linear public goods games have incentives that make them resemble a many-person PD game. Individuals have an endowment m which they each must allocate between themselves and a public account. Each of the n members of the group earns α for each dollar allocated to the public account. By design, $0 < \alpha < 1$, so giving nothing is a dominant strategy, but $\alpha n > 1$, so giving m is Pareto efficient.

The results of these games are that average giving is significantly above zero, even as we change n, m and α (Isaac and Walker 1988; Isaac et al. 1994) and whether the play is with the same group of 'partners' or with randomly changing groups of 'strangers' (Andreoni 1988). Hence,

reputations play little role in public goods games (Andreoni and Croson 2008; Palfrey and Prisbrey 1996).

In his review of this literature, Ledyard (1995) notes that, with a dominant strategy of giving zero, any error or variance in the data could mistakenly be viewed as altruism. Thus, to determine what drives giving one needs to confirm that subjects understand the dominant strategy but choose to give anyway.

Andreoni (1995) develops a design to separate 'kindness' from 'confusion' in linear public goods games. Rather than paying subjects for their absolute performance, in one treatment he paid subjects by their relative performance. Converting subjects' ranks into their payoffs converts a positive-sum game to a zerosum game. It follows that even altruists have no incentive to cooperate when paid by rank (that is, under the usual definition of altruism where people love themselves at least as much as they love others). Cooperation by subjects in the treatment group, therefore, provides a measure of confusion. Andreoni finds that both kindness and confusion are significant, and about half all cooperation in public goods games is from people who understand free riding but choose to give anyway.

To establish that giving is deliberate, however, does not necessarily mean it is based in altruism; it could, instead, be from warm-glow. Two papers, using similar experimental designs but different data analysis methods, explore this question by separating the marginal net return that a gift to the public good has for the giver and for the recipient. The 'internal return' experienced by the giver should affect warm-glow and altruism, but the 'external return' received by the others affects only altruism. Palfrey and Prisbrey (1997) find that warm-glow dominates altruism, while Goeree et al. (2002) find mostly altruism. Combining this evidence, it appears that both motives are likely to be significant.

Another way to test for the presence of altruism and warm-glow is to choose a manipulation that would have different predictions in the two regimes. Andreoni (1993) looks at the complete crowding out hypothesis, which states that a lump-sum tax, used to increase government spending on a public good, will reduce an altruist's voluntary contributions by the amount of the tax. He employs a public goods game with an interior Nash equilibrium. Suppose subjects care only about the payoffs of other subjects (altruism). Then if we force subjects to make a minimum contribution below the Nash equilibrium, this should simply crowd out their chosen gift, leaving the total gift unchanged. If they get utility from the act of giving (warm-glow), by contrast, crowding out should be incomplete. Andreoni finds crowding at 85 per cent, which is significantly different from both zero and 100 percent. This confirms the findings from the last paragraph; both warm-glow and altruism are evident in experiments on public goods. Similar findings are presented in Bolton and Katok (1998) and Eckel et al. (2005).

Dictator Games

This line of research began with the ultimatum game, where a proposer makes an offer on the split of a sum of money. If the responder accepts, the offer is implemented, while if she rejects both sides get nothing. Guth et al. (1982) find that proposers strike fair deals and leave money on the table. Is this altruism, or just fear of rejection? To answer this question Forsythe et al. (1994) also examine behaviour in a dictator game that cuts out the second stage, leaving selfish proposers free to keep the whole pie for themselves, and leaving altruists unconstrained to give a little or a lot. While keeping the entire endowment is the modal choice in the dictator game, a significant fraction of people give money away. On average, people share about 25 per cent of their endowment. This seems to indicate significant altruism.

Again, researchers have explored numerous non-altruistic explanations. One is that, while the dictator's identity is unknown to the recipient, it is not unknown to the researcher. This lack of 'social distance' could cause the selfish but selfconscious subjects to give when they would prefer not to. Hoffman et al. (1994) take elaborate steps to increase the anonymity and confidentiality of the subjects so that even the researcher cannot know their choices for sure. They find that this decreases giving to about 10 per cent of endowments.

However, this 'double anonymous' methodology creates problems of its own. Bolton et al. (1998) argue that greater anonymity makes the participants sceptical about whether the transfers will be carried out. Bohnet and Frey (1999) find that reducing the social distance increases equal splits greatly, but in their anonymous treatments giving again averages 25 per cent (see also Rege and Telle 2004).

Andreoni and Miller (2002) take a different approach. They note that, if altruism is a deliberate choice, then it should follow the neoclassical principles of revealed preference. They gave subjects a menu of several dictator 'budgets', each with different 'incomes' and different 'prices' of transferring this income to another anonymous subject. By checking choices against the generalized axiom of revealed preference, they show that indeed most subjects are rational altruists, that is, they have consistent and well-behaved preferences for altruistic giving in a dictator game. They also show substantial heterogeneity across subjects, with preferences ranging from utilitarian (maximizing total payments to both subjects) to Rawlsian (equalizing payments to both subjects). Interestingly, men and women are on average equally altruistic in this study, but vary significantly in response to price. Andreoni and Vesterlund (2001) show that men are more likely to be utilitarian, and women are more likely to be Rawlsian. This implies that men are significantly more generous when giving is cheap (that is, it costs the giver less than one to give one), but women are significantly more altruistic when giving is expensive (costs greater than or equal one to give one). Which is the fairer sex, therefore, depends on the price of giving (see also Eckel and Grossman 1998, on dictator games when the price is one).

Trust Games and Gift Exchange

When someone buys a loaf of bread from a baker, there is a moment when one party has both the bread and the money and the incentive to take both. Why don't they? Similarly, why are some car mechanics truthful, and why do some workers put in an honest effort even when they are not monitored? These questions have been studied under names of trust games and gift exchange.

In the trust game, two players are endowed with M each. A sender chooses to pass x to a receiver. A receiver receives kx, where $k > 1$. The receiver then chooses a y to pass back to the sender. Senders earn $M - x + y$, while receivers earn $M + kx - y$. Since $y = 0$ is a dominant strategy for receivers, $x = 0$ is the subgame perfect equilibrium strategy for senders. That is, since the baker keeps both the bread and the money, no exchange is attempted. Despite this dire prediction, x and y are often positive, and y is typically increasing in x. While there is tremendous variance, the average y is often slightly below the average x (Berg et al. 1995).

The gift exchange game is a nonlinear version of the trust game above. Fehr et al. (1993) adapted the Akerlof (1982) labour market model of efficiency wages. Some subjects play the roles of firms and offer labour contracts to workers. The contracts stipulate a wage and an expected effort level of workers. Since effort is costly and unobservable, it should be minimal. The subjects playing the role of firms should expect low effort, and offer low wages. However, in the experiment wages are high and effort rises with the wage offer, just as Akerlof predicted.

Trust and gift exchange games are often used to argue for the importance of reciprocity. Reciprocity is, however, an ulterior motive – giving in order to either generate or relieve an obligation is not altruism by the definition in our introduction. How much of the exchange can be attributed to altruism alone? Cox (2004) separates these motives by comparing senders in a trust game with those in a dictator game. As dictators have no ulterior motive of generating an obligation, their behaviour can be used to estimate the altruism of senders. For receivers he uses a control group whose x is determined at random by the experimenter. These receivers have no obligation to the sender, thus their transfers serve as a measure of the receivers' altruism. Cox finds that 60 per cent of an average sender's x and 42 per cent of the average receiver's y is motivated by altruism. Thus, while reciprocity is clearly present, altruism is not replaced in this exchange (see also Charness and Haruvy 2002; Gneezy et al. 2000).

While some have criticized whether gift exchange in the laboratory is robust to small changes in parameters and presentation (Charness et al. 2004), others have challenged gift exchange in the field. List (2006) looks for gift exchange on the trading floor of a sports card market. He conducts a series of experiments that move incrementally from a standard laboratory game with a neutral presentation to actual exchanges on the floor. While he finds that gift exchange (higher-quality product in return for higher price) is not totally extinguished in the actual market, he also finds that reputation is far more important in determining the quality provided by sellers. Gneezy and List (2006) follow up with a labour market experiment. They recruited students to do a one-day job working in a library. The treatment group was told, unexpectedly, that their wage would be 167 per cent of the agreed wage. These subjects were significantly more productive in the first 90 minutes of work than the control subjects. However, after a one-hour lunch break, there was no difference between the productivity of treatment and control. They conclude that gift exchange in actual labour markets may have no long-term effects.

Conclusion

There is ample consistent evidence of altruism in experiments. This follows both from studies that have taken great effort to remove any ulterior motives, as well as studies that provide manipulations that should influence altruism. While the existence and importance of altruism seem well established in the laboratory, many questions that could help us understand and amplify altruism remain unanswered.

First, where do altruistic preferences come from? One notion is that they come from culture. Evidence of this is suggested by differences in behaviour in experiments in different countries (Roth et al. 1991; Henrich et al. 2001). Another notion is that they are acquired as part of psychological development and socialization, as seen in economic experiments using children as subjects (Harbaugh and Krause 2000). A third possibility for altruism is that we are innately wired to care. Harbaugh et al. (2007) use fMRI to show that neural activation in the ventral striatum is very similar when money goes to the subject and when it goes to a charity, and that the relative activations actually predict who will give. Tankersley et al. (2007) show that posterior superior temporal sulcus activation is higher for people who report more helping behaviour outside the lab.

Second, is altruism significant outside the laboratory? The laboratory is, after all, a unique environment. Field experiments on fundraising, such as List and Lucking-Reiley (2002), show the potential of this method for finding good evidence of altruism outside the laboratory, but without giving up all experimental control.

Finally, how does altruism combine with other ulterior motives? Are warm-glow and altruism inextricably linked, and can we use mechanisms that act on warm-glow to amplify altruism and overcome free riding? Does voting to force everyone to provide a public good provide a warm-glow benefit to the voters? Economic experiments may be a productive method for answering these questions, and for using the knowledge of altruism that results to improve the institutions within which altruist economic agents interact.

See Also

- ▶ Altruism, History of the Concept
- ▶ Charitable Giving
- ▶ Experimental Economics
- ▶ Experimental Economics, History of
- ▶ Public Goods
- ▶ Public Goods Experiments

Bibliography

Akerlof, G.A. 1982. Labor contracts as partial gift exchange. *Quarterly Journal of Economics* 97: 543–569.

Andreoni, J. 1988. Why free ride? Strategies and learning in public goods experiments. *Journal of Public Economics* 37: 291–304.

Andreoni, J. 1989. Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy* 97: 1447–1458.

Andreoni, J. 1990. Impure altruism and donations to public goods: A theory of warm-glow giving. *Economic Journal* 100: 464–477.

Andreoni, J. 1993. An experimental test of the public-goods crowding-out hypothesis. *American Economic Review* 83: 1317–1327.

Andreoni, J. 1995. Cooperation in public-goods experiments: Kindness or confusion? *American Economic Review* 85: 891–904.

Andreoni, J., and R. Croson. 2008. Partners versus strangers: The effect of random rematching in public goods experiments. In *Handbook of experimental economics results*, ed. V. Smith and C. Plott. New York: North-Holland, forthcoming.

Andreoni, J., and J.H. Miller. 1993. Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *Economic Journal* 103: 570–585.

Andreoni, J., and J.H. Miller. 2002. Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* 70: 737–753.

Andreoni, J., and L. Samuelson. 2006. Building rational cooperation. *Journal of Economic Theory* 127: 117–154.

Andreoni, J., and L. Vesterlund. 2001. Which is the fair sex? Gender differences in altruism. *Quarterly Journal of Economics* 116: 293–312.

Berg, J., J. Dickhaut, and K. McCabe. 1995. Trust, reciprocity, and social history. *Games and Economic Behavior* 10: 122–142.

Bohnet, I., and B.S. Frey. 1999. Social distance and other-regarding behavior in dictator games: Comment. *American Economic Review* 89: 335–339.

Bolton, G.E., and E. Katok. 1998. An experimental test of the crowding out hypothesis: The nature of beneficent behavior. *Journal of Economic Behavior and Organization* 37: 315–331.

Bolton, G.E., E. Katok, and R. Zwick. 1998. Dictator game giving: Rules of fairness versus acts of kindness. *International Journal of Game Theory* 27: 269–299.

Camerer, C., and K. Weigelt. 1988. Experimental tests of a sequential equilibrium reputation model. *Econometrica* 56: 1–36.

Charness, G., G.R. Frechette, and J.H. Kagel. 2004. How robust is laboratory gift exchange? *Experimental Economics* 7: 189–205.

Charness, G., and E. Haruvy. 2002. Altruism, equity, and reciprocity in a gift-exchange experiment: An encompassing approach. *Games and Economic Behavior* 40: 203–231.

Cox, J.C. 2004. How to identify trust and reciprocity. *Games and Economic Behavior* 46: 260–281.

Eckel, C.C., and P.J. Grossman. 1998. Are women less selfish than men? Evidence from dictator experiments. *Economic Journal* 108: 726–735.

Eckel, C.C., P.J. Grossman, and R.M. Johnston. 2005. An experimental test of the crowding out hypothesis. *Journal of Public Economics* 89: 1543–1560.

Fehr, E., G. Kirchsteiger, and A. Riedl. 1993. Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics* 108: 437–459.

Forsythe, R., J.L. Horowitz, N.E. Savin, and M. Sefton. 1994. Fairness in simple bargaining experiments. *Games and Economic Behavior* 6: 347–369.

Gneezy, U., W. Guth, and F. Verboven. 2000. Presents or investments? An experimental analysis. *Journal of Economic Psychology* 21: 481–493.

Gneezy, U., and J.A. List. 2006. Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica* 74: 1365–1384.

Goeree, J.K., C.A. Holt, and S.K. Laury. 2002. Private costs and public benefits: Unraveling the effects of altruism and noisy behavior. *Journal of Public Economics* 83: 255–276.

Guth, W., R. Schmittberger, and B. Schwarze. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* 3: 367–388.

Harbaugh, W.T., U. Mayr, and D. Burghart. 2007. Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* 316: 1622–1625 .forthcoming

Harbaugh, W.T., and K. Krause. 2000. Children's altruism in public good and dictator experiments. *Economic Inquiry* 38: 95–109.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath. 2001. In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review* 91: 73–78.

Hoffman, E., K. McCabe, K. Shachat, and V.L. Smith. 1994. Preferences, property rights and anonymity in bargaining games. *Games and Economic Behavior* 7: 346–380.

Isaac, M.R., and J.M. Walker. 1988. Group size effects in public goods provision: The voluntary contributions mechanism. *Quarterly Journal of Economics* 103: 179–199.

Isaac, M.R., J.M. Walker, and A.W. Williams. 1994. Group size and the voluntary provision of public goods: Experimental evidence utilizing large groups. *Journal of Public Economics* 54: 1–36.

Kelley, H.H., and A.J. Stanelski. 1970. Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology* 16: 66–91.

Kreps, D., P. Milgorm, J. Roberts, and R. Wilson. 1982. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory* 27: 245–252.

Ledyard, J.O. 1995. Public goods: A survey of experimental research. In *The handbook of experimental economics*, ed. J.H. Kagel and A.E. Roth. Princeton: Princeton University Press.

List, J.A. 2006. The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *Journal of Political-Economy* 114: 1–37.

List, J.A., and D. Lucking-Reiley. 2002. The effects of seed money and refunds on charitable giving: Experimental evidence from a university capital campaign. *Journal of Political Economy* 110: 215–233.
McKelvey, R.D., and T.R. Palfrey. 1992. An experimental study of the centipede game. *Econometrica* 60: 803–836.
Nagel, T. 1970. *The possibility of altruism*. Oxford: Clarendon Press.
Palfrey, T.R., and J.E. Prisbrey. 1996. Altruism, reputation and noise in linear public goods experiments. *Journal of Public Economics* 61: 409–427.
Palfrey, T.R., and J.E. Prisbrey. 1997. Anomalous behavior in public goods experiments: How much and why? *American Economic Review* 87: 829–846.
Rege, M., and K. Telle. 2004. The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics* 88: 1625–1644.
Roth, A.E., and J.K. Murnigham. 1978. Equilibrium behavior and repeated play of the prisoner's dilemma. *Journal of Mathematical Psychology* 17: 189–198.
Roth, A.E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir. 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review* 81: 1068–1095.
Tankersley, D., C.J. Stowe, and S.A. Huettel. 2007. Altruism is associated with an increased neural response to agency. *Nature Neuroscience* 10: 150–151.

Altruism, History of the Concept

Philippe Fontaine

Abstract

This article describes the incorporation from the early 1960s of seemingly unselfish behaviour into economics. Faced with the problem of accounting for such behaviour in a discipline that often relies on the selfishness assumption, some economists used the notion of sympathetic preferences within the self-interest model, whereas others tried instead to supplement that model with an ethically inspired model. It is unclear that in investigating seemingly unselfish behaviour, economists have gained a better understanding of its actual motivations, but in the process they have been led to take more seriously other conceptions of human being than economic man.

Keywords

Altruism; Becker, G.; Biology and economics; Boulding, K.E.; Buchanan, J.M.; Charitable giving; Choice; Commitment; Economic man; Ethics and economics; Family decision making; family economics; Free-rider problem; *Homo economicus*; Institute of Economic Affairs; Non-economic behaviour; Phelps, E.; Philanthropy; Preference; Reciprocity; Ricardian equivalence theorem; Robertson, D.; Rotten kid theorem; Self-interest; Sen, A.; Sociobiology; Titmuss, R.M.; Utility interdependence; Vickrey, W.S.; Warm-glow hypothesis

JEL Classifications

B0

Dennis Robertson once asked: 'What does the economist economize?' (1955, p. 154). His answer was: '[T]hat scarce resource Love – which we know, just as well as anybody else, to be the most precious thing in the world'. He meant that a better understanding of the economy had the happy consequence of allowing people to conduct their business without having to rely excessively on social virtues. For upholders of economic man, that was certainly a good justification for doing without 'love'. And had it not been for a study of philanthropy, conducted in the late 1950s, they might well have continued to ignore 'love of human kind in general', as dictionaries usually define it.

The reintroduction of what is regarded today as 'altruism' into contemporary economics, following Edgeworth's (1881, p. 53n) first modern formulation in the late 19th century, came out of this effort to understand philanthropy, and not, as conventional wisdom suggests, from the publication of Gary Becker's (1974) 'A Theory of Social Interactions', which was but one tardy sequel to it. In writing a history of recent work on unselfishness, therefore, it is crucial that Becker's two chief results in that article, namely, the invariance proposition and the '"rotten kid" theorem', do not mask the sheer diversity of research before the

mid-1970s nor the tensions that persisted afterwards. In what follows, we describe the key moments that preceded the inclusion of an 'altruism' heading in the *Journal of Economic Literature* (JEL) classification system for journal articles at the end of 1993.

Understanding Philanthropy

After private foundations fell under increasing regulatory scrutiny in the early 1950s (Hall 1999) and their tax status was attacked in the early 1960s (Frumkin 1999), their leaders found it opportune to approach economists for advice. In effect, Donald Young, President of the Russell Sage Foundation (RSF), asked Solomon Fabricant, Director of Research at the National Bureau of Economic Research (NBER), to think about the possibility of investigation into the economic aspects of philanthropy. The RSF eventually funded a study of this phenomenon in the American economy, which was conducted by the NBER between 1959 and 1962 under the supervision of the economist Frank G. Dickinson, who was assisted by an advisory committee. The first meeting of the committee took place in late 1959.

A few members of the NBER staff, notably Becker, attended the meeting. Some work stemming from it, notably by Fabricant and Dickinson, and dealing mostly with definitional and empirical aspects, benefited from a limited circulation. That explains why Becker wrote the obscure and unpublished 'Notes on an Economic Analysis of Philanthropy' in April 1961. He later identified this article as the first expression of his interest in social interactions, but it was originally just another effort to extend the utility maximization assumption to the study of 'non-economic' topics. Becker's 'Notes' was not the only outgrowth of the NBER project. In addition, a conference, envisioned by Dickinson, took place in June 1961, bringing together a number of economists, among whom William Vickrey and Kenneth Boulding gave papers and James Buchanan simply attended.

By the early 1960s, then, the study of philanthropy had provided an opportunity for a handful of economists to explore aspects of seemingly unselfish behaviour. Following Dickinson's remark, in late 1959, that philanthropy was not in the mainstream of economic analysis, Becker (1961), Vickrey (1962) and Boulding (1962) suggested that there were no theoretical impediments to its understanding. 'It can be dealt with quite easily in utility theory', wrote Boulding, 'by considering the utility of one person a function not only of his own wealth or his own income, but a function of the wealth and income of others' (1962, p. 61). Essentially Becker and Vickrey agreed. Utility interdependence, in its modern form, had long been around and it appeared to be the proper tool to tackle philanthropic behaviour, even if there could be variations in the arguments to be included in the giver's utility function. There was, however, a more significant difference. Becker was not especially concerned with the motivations of philanthropic behaviour, whereas Boulding and Vickrey were: they believed utility theory could not elucidate the variety of motivations for philanthropy. Accordingly, Boulding emphasized the sense of community (and the associated capacity for empathy) as the essence of 'genuine philanthropy', while Vickrey saw social distance as the significant factor.

Following the work of Becker, Vickrey and Boulding, various research efforts gave momentum to the study of philanthropy. Interested as he was in the effects of fiscal systems on income redistribution, Buchanan could easily relate to the theme of the philanthropy conference. As was the case for Becker and Boulding, his work at the intersection of economics and other social sciences made the whole undertaking of studying a form of seemingly unselfish behaviour especially appealing to him. His own views on the free-rider problem led him to distinguish between the expediency criterion and moral law as the two main determinants of an individual's choice, and to connect their relative strength to group size (Buchanan 1965). The individual was said to follow moral law in small-group interactions and turn into a utility maximizer as soon as group size grew – the 'large-group ethical dilemma'. In his presidential address to the American

Economic Association a few years later, in December 1968 when social crisis was at its height, Boulding (1969), too, felt it timely to contrast two sets of common values guiding human behaviour: the 'economic ethic' and the 'heroic ethic', with the former centring on cost-benefit analysis and the latter emphasizing the sense of identity.

After his resignation from the University of Virginia in 1968, Buchanan visited UCLA where a number of economists, including Armen Alchian and William Allen (1964) and Jack Hirshleifer (1967), had studied seemingly unselfish behaviour. These authors had reached the conclusion that its treatment did not require a fundamental reconsideration of the behavioural assumptions of economic theory. Buchanan had doubts. Although he recognized the merits of enriched utility functions for the study of seemingly unselfish behaviour, Buchanan warned that they did not unveil the variety of human motivations. Consequently, he argued, the inclusion of 'noneconomic' arguments, such as love or concern for the welfare of others, into the utility function did not necessarily improve the predictive power of theory.

Understanding Altruism

In the context of adverse circumstances for foundations, the 1962 conference was meant to correct the inadequacy of knowledge about the economic aspects of philanthropy. In 1971, Edmund Phelps sent a grant proposal to Orville Brim, Jr., then President of the RSF, to ask support for the organization of a conference to be held in New York City. Here, too, the idea of the conference emerged in a difficult political environment. With the Tax Reform Act of 1969 imposing new regulations, foundations leaders were under pressure to defend philanthropy from any further threat. Unlike its predecessor, however, the conference contemplated by Phelps would not deal with an instance of seemingly unselfish behaviour, but with altruistic behaviour in general.

Pointing to the extension of the domain of economics to neglected topics such as crime and war, to the disenchantment with classical liberalism that accompanied the intensification of economic problems and the deepening of social crisis in the United States, and to new developments in the analysis of markets such as the relaxation of the assumption of perfect information, Phelps concluded: 'the time has arrived for a theory of altruism' (Phelps to Brim, 19 October 1971). That the conference was meant to deal with a topic, the definition of which was still unclear to many, including Phelps himself, speaks volumes about the appeal of seemingly unselfish behaviour in social science at a time when 'the amount of divisiveness and conflict in a society' – to use Mancur Olson's (1971, p. 173) words – occasioned serious concern.

Phelps's consideration of possible participants reveals that what the profession has come to call 'altruism' was in the early 1970s a heterogeneous body of knowledge comprising disparate analyses of human behaviour. Phelps first contacted Kenneth Arrow, Paul Samuelson and Vickrey, who all agreed to present papers. In his proposal, Phelps mentioned Boulding, Thomas Schelling, Becker, James Mirrlees, Peter Hammond, Sydney Winter, Alchian, Duncan Foley and Scott Boorman. Among non-economists, philosophers had the lion's share in an otherwise odd group including John Rawls, Tom Nagel, Marshall Cohen, Erving Goffman, Edward Banfield, Bernhard Lieverman and Sydney Morgenbesser. Several of these researchers were part of a movement in the late 1960 and early 1970s to connect moral philosophy with economics and other social sciences. And many of them were concerned with the respective role of self-interest and ethics in the explanation of human behaviour.

Amartya Sen did not appear in the list above but he attended the conference. His call for reconsidering the economic theory of human behaviour fitted in well with the overall preoccupation of the conference with ethics. In his LSE inaugural lecture, 'Behaviour and the Concept of Preference', Sen (1973) offered valuable insights into the relationships between choices and individual preferences, showing that the same choice (use and reuse of glass bottles) could correspond to four distinct cases

in terms of the agent's underlying preferences. The first three cases represented the preferences of a selfish, sympathetic and socially conscious individual, respectively; they were consistent with utility theory. The fourth case, which Sen (1977) later associated with the notion of 'commitment', was of a different sort, however. It shows that moral considerations could influence individual choice in such a way as to undermine the correspondence between choice and preference on the one hand and preference and welfare on the other. The maximization framework with utility interdependence told some truth about seemingly unselfish behaviour, but not the only truth.

However, not all students of seemingly unselfish behaviour found ethics illuminating. At about the same time as Sen's LSE lecture was published in August 1973, Arthur Seldon, from the Institute of Economic Affairs (IEA), the London-based think-tank, was completing the preface to *The Economics of Charity: Essays on the Comparative Economics and Ethics of Giving and Selling, with Applications to Blood.* Unlike Sen and others, the main contributors to the collection, including Alchian, Allen and Gordon Tullock, were doubtful about the possibility of learning something significant economically from an ethical approach to unselfish behaviour. They preferred instead to explore the potentialities of utility theory.

In the early 1970s, economists were undoubtedly showing greater interest in what was now occasionally called 'altruism', but a unified theory was still lacking. The plurality of viewpoints reflected varied motivations, with some striving to renew the understanding of small-group interactions and others discussing either the moral dimension of economic behaviour or the economic dimension of moral behaviour. In the literature, there emerged a dividing line between the advocates of *homo economicus* and the supporters of *homo ethicus,* which became more pronounced with the publication of Becker's 'A Theory of Social Interactions' (1974) and Phelps's *Altruism, Morality, and Economic Theory* (1975), a collection of essays resulting from the New York conference.

The Polarization of the Mid-1970s

Becker's 1974 article was originally titled 'Interdependent preferences: charity, externalities and income taxation': it was renamed in September 1969 – a change that revealed Becker's intention to broaden his frame of analysis from the issue of charity to the treatment of seemingly unselfish behaviour in general. The article was published in the same issue of the *Journal of Political Economy* as Robert Barro's 'Are Government Bonds Net Wealth?' (1974). It would be unreasonable to think of Barro's analysis of government budget deficits as a simple application of Becker's 'rotten kid' statement, but some cross-fertilization occurred, especially since Becker's manuscript had spent some six years in his files and Barro had commented on it. Becker also knew Barro's article, a draft of which had been presented, in 1973, in the Money and Banking workshop run by Milton Friedman in Chicago. That Becker and Barro discussed seemingly unselfish behaviour is evidenced by the fact that the latter's former wife suggested the phrase 'rotten son' to the former who later turned it into 'rotten kid' in his eponymous 'theorem' (Barro to Fontaine, 3 April 2001, personal communication).

Becker proposed, in contrast to what he called the 'usual theory of consumer choice', which places in the utility function of the giver his own consumption together with the amount of his charitable giving, a 'social interactions' approach, which replaces the amount of charitable giving with the consumption of beneficiaries, as financed by their income and the amount of charitable giving they receive. In the context of the family, Becker reached the conclusion 'that if a [benevolent] head exists, *others members also are motivated to maximize family income and consumption, even if their welfare depends on their own consumption alone*. This is the '"rotten kid" theorem' (1974, p. 1080).

Against the background of a family break-up, Becker showed that the conditions for family cohesion were not so demanding as to require that all family members have sympathetic preferences or so unrealistic as to imply that all family members are selfish. Regarding the recipients of

the head's generosity, he endorsed Friedman's (1953) influential argument and made it clear that only '*as-if* altruism' was involved. Yet the head had sympathetic ('altruistic') preferences. In other words, his transfers were said to result from sympathy, which was explained by the fact that 'the marriage market is more likely to pair a person with someone he cares about than with an otherwise similar person that he does not care about' (Becker 1974, p. 1074n).

Assuming continuity between family and other groups, Becker extended his results to the 'synthetic "family"', consisting of a charitable person and all recipients of his or her charity, and to a number of other multi-person interactions. Here again, due to offsetting changes in transfers from the sympathetic benefactor, a redistribution of income among 'members' left their own welfare unchanged. To the problem represented by the possibility that opportunistic tendencies can surface in groups characterized by the interactions of selfish individuals and therefore prevent socially desirable outcomes, in the mid-1970s Becker offered a solution centred on the sympathetic preferences of certain individuals in society. To many today, this answer will seem ad hoc, but, at a time when much was said about the unresponsiveness of people to each other's lot, it went against the stream. With the increase in macroeconomic volatility, its policy implications were, however, straightforward: due to offsetting private transfers, one could hardly count on social and economic policies to change the distribution of resources (see Barro 1974).

Though it can be argued that 'A Theory of Social Interactions' played a significant role in the history of unselfishness research, it should be remembered that its main objective was to analyse the economic implications of interactions within various groups. Phelps (1975), by contrast, meant to offer a contribution to the 'theory of altruism'. As such he aimed at understanding a variety of behaviours, the motivations of which were seemingly unselfish. While Becker had provided a coherent framework centred on maximization with utility interdependence to analyse social interactions, the essays in Phelps's collection illustrated the complexity, indeed vagueness, of 'altruism' as soon as one ventures beyond the self-interest model.

In dealing with unselfishness, Phelps's book actually considered a great variety of behaviours and motivations. Accordingly, contributors strove to classify them so as to identify their similarities and differences. When Arrow (1975) discussed Richard Titmuss's analysis of blood giving and its motives, for instance, he introduced a distinction between benefiting from the satisfactions obtained by others, benefiting from one's contributions to these satisfactions and the idea that 'each performs duties for the other in a way calculated to enhance the satisfaction of all' (1975, p. 17), but he refrained from providing an economic translation of Titmuss's reference to a sense of obligation to strangers. Arrow acknowledged the possibility that individuals act according to a categorical imperative, but noted: 'I should add that, like many economists, I do not want to rely too heavily on substituting ethics for self-interest' (1975, p. 22).

Others in the volume were probably more willing to take note of ethical motivations if only because they could serve to justify opposition to governmental regulation in various areas. In 'The Samaritan's Dilemma', Buchanan (1975) showed that the expectation of other-oriented behaviour could lead the potential beneficiary to behave opportunistically. Of particular interest in Buchanan's approach was the association of the undesirable consequences of other-oriented behaviour with the prevalence of the expediency criterion (the selfishness of agents) in society and the conclusion that commitment à la Schelling offered a solution to that problem. This solution, Buchanan realized, was threatened by the weakening adherence to ethical rules resulting from increase in group size.

The last three essays in Phelps's volume came back to the issue of philanthropy. Of particular interest was Bruce Bolnick's (1975) acknowledgement that a number of writers had 'rendered such behavior susceptible to the traditional tools of economic analysis' and his concomitant remark that 'a more fundamental issue is uncovered: What types of motivation underlie philanthropic activity?' (1975, p. 197). In the same vein, Bolnick pointed to the difference between trying to understand seemingly unselfish behaviour

and studying the consequences of the inclusion of utility interdependence in the maximization framework in terms of optimality conditions (see, for example, Hochman and Rodgers 1969; Kolm 1969; Thurow 1971). The latter approach Bolnick saw as 'unsatisfying as a behavioral theory' (1975, p. 198) and accordingly argued that social rewards and psychological consistency had to be taken into account not only for small groups but also for larger ones. In the process, Bolnick mentioned the justification in terms of empathetic identification, as suggested by Boulding (1962) and Vickrey (1962), but expressed uneasiness with its limitation to close-knit groups.

Despite notable efforts to go beyond the self-interest model, *Altruism, Morality, and Economic Theory* failed to identify the main features of the 'commitment model'. The fact that ethical considerations had to be taken into account in the analysis of seemingly unselfish behaviour did not mean that the self-interest model failed on most accounts or that another model could claim greater explanatory power. It is understandable therefore that in his Introduction to the volume Phelps (1975) wavered: 'Can altruistic behavior be fit into some version of the economist's beloved model of utility maximization subject to constraints? Or must that model be importantly modified and hooked up to some complementary body of analysis to yield a satisfactory product?' (1975, p. 2). Jean-Jacques Laffont (1975) conveyed some of these tensions when he uncharacteristically defined the behaviour of *homo economicus* as selfish, not self-interested, and contrasted it with 'Kantian' behaviour.

The Self-interest View of Unselfishness

With the studies of seemingly unselfish behaviour within the framework of utility maximization with interdependence, the question of the arguments to be included in the utility function became more relevant than that of the actual motivations of behaviour, though these arguments have occasionally been equated with motives for action.

The malleability of utility functions made it possible for economists to consider a variety of influences on the satisfaction of the individual besides own consumption. It even allowed for the inclusion of biological arguments into the utility function. Becker's (1976a) review article on Edward Wilson's (1975) controversial *Sociobiology* provides an interesting illustration. To Wilson, who suggested that biology might enlighten the analysts of social behaviour, Becker, who by that time saw himself as one of them, replied that economics too had its merits in terms of explaining the 'social' (for illustrations of the 'economic approach', see Becker 1976b). Thus, though Becker accepted Wilson's definition of *altruism* as behaviour that reduces one's genetic fitness to the benefit of another's, he also pointed out that 'altruism', because of its effects on the behaviour of beneficiaries, could increase the genetic fitness of the 'altruist'. In emphasizing the positive outcome of unselfish behaviour for the 'altruist', Becker complicated the emerging discourse on the essentially selfish nature of human behaviour, as derived from the view that 'altruism' is detrimental to its author (Dawkins 1976).

There was indeed something accidental about Becker's considering the biological basis of social behaviour and writing about sociobiology, but for economists taken by the 'economic approach' there was good reason to address unselfishness: economics could not hope to embrace anthropological, sociological and political subjects without at the same time breaking away from the advocacy of behavioural assumptions that pictured the economic agent as a non-social being.

The second half of the 1970s offered several examples of authors, among whom were Hirshleifer and Tullock, who advocated the expansion of the 'economic' and wrote on unselfishness as well. It is hardly surprising therefore that these two commented on Becker's article in the *Journal of Economic Literature*. Hirshleifer (1977a), whose extremely well-documented 'Economics from a Biological Viewpoint' had just appeared in the *Journal of Law and Economics*, another symbol of the expansionist ambitions of economics, noted that the '"rotten kid" theorem' obtained only if the 'altruistic' head had the last word in the decision sequence (Hirshleifer to

Becker, 13 December 1976; see also Hirshleifer 1977b). Hirshleifer's proviso suggested paradoxical implications. If the 'head' did not have the last word, the theorem lost its strength as a demonstration that selfish individuals were dissuaded from behaving opportunistically in groups; if he or she did, on the other hand, it might be presumed that some of the problems dealt with in the 'theorem' lost significance.

Unlike Becker, Tullock (1977) preferred a model of unselfishness in which the giver derives utility from the mere act of giving. In his comment, he made the interesting point that in Becker's model the giver does not necessarily know the preference ordering of recipients. Becker thought this problem irrelevant since his model was concerned with family, not government, transfers (Becker to Tullock, 14 December 1976). Such a justification, it should be noted, could undermine the claim that his argument reached beyond the kin selection explanation of unselfishness by biologists.

As Hirshleifer's and Tullock's reactions to Becker's inroad into sociobiology illustrate, some economists were interested in biology. Though the impetus came from the heated debates surrounding the publication of *Sociobiology*, the ongoing redefinition of territories in social science was the determining factor. In his review of the literature on the relationships between economics and biology, Hirshleifer (1977a) noted that 'the social sciences generally can be regarded as in the process of coalescing' (1977a, p. 3) and he concluded that 'economics can be regarded as the general field, whose two great subdivisions consist of the natural economy studied by the biologists and the political economy studied by economists proper' (1977a, p. 52). Clearly, economists were unwilling to see their attempts at investigating the 'social' threatened by similar ambitions on the side of natural scientists (see Hirshleifer 1985, who later spoke of 'competing imperialisms' but acknowledged their complementarities), especially since these attempts continued to be regarded suspiciously by some in the profession. Accordingly, economists took every occasion to emphasize economics' lessons for the natural sciences. Becker did this and so did others, including Boulding (1978), Hirshleifer (1977a), Schelling (1978), Tullock (1978, 1979), who all took an interest in studying 'non-economic' behaviour.

Though these various initiatives enjoyed greater visibility with the organization of a session on 'Economics and Biology: Evolution, Selection, and the Economic Principle' at the meeting of the American Economic Association in December 1977, from the early 1980s unselfishness research was conducted independently of sociobiology. With 'economics imperialism' gradually entering the mainstream (see Stigler 1984; Hirshleifer 1985), the interest of economists turned to the more general study of the relationships between economics and biology (see for example, Hirshleifer 1982; Nelson and Winter 1982; Samuelson 1985), and it is only in the early 1990s that the question of unselfishness surfaced again in this kind of literature (Tullock 1990; Simon 1990, 1992, 1993; Bergstrom and Stark 1993; Samuelson 1993).

By the early 1980s, the self-interest view of unselfishness was well established in the profession: it associated 'altruism' with the fact that an individual's utility function depended on another's well-being. Becker's (1981, p. 2) 'Altruism in the Family and Selfishness in the Market Place' illustrated the main orientations of that view when he noted that his was a definition of *altruism* that concerned behaviour, not 'a philosophical discussion of what "really" motivates people', and that 'altruism' was more common in the family than in the market place because of its greater relative efficiency in the former (1981, p. 10).

The departures from 'altruism' à la Becker were encouraged by the political debates of the 1980s. With the macroeconomic volatility of the 1970s, the bearing of economics on policy matters began to be challenged. The conclusion, that due to offsetting transfers from 'altruists' one could hardly count on social and economic policies to change the distribution of resources, found continuation in various remarks about the 'ungovernability' of modern societies (see Olson 1982, p. 8). And with the beginning of Ronald Reagan's first presidency and its economic programme turning away from demand management, the link between the ineffectiveness

of governmental redistribution and the existence of sympathetic transfers took up a broader significance; it could be taken as another argument for lesser state intervention.

With significant changes in economic and social policies looming on the horizon in the first half of the 1980s, notably 'the control of federal spending, the reduction or elimination of a wide variety of social entitlement and redistributive schemes … and the aggressive reduction of tax rates on incomes' (Bernstein 2001, p. 164), a number of economists were led to re-examine the strength of the '"Ricardian equivalence" theorem' and the ' "rotten kid" theorem', two results that were closely associated with the unselfishness literature.

Becker's (1981) article appeared in February at a time when President Reagan's programme was being presented. That programme carried with it a vision of the workings of society that some of Reagan's predecessors considered mistaken, precisely because it gave inadequate weight to the failures of the invisible hand of the market. In the 1960s and 1970s, some may have thought seemingly unselfish behaviour a solution to the opportunistic tendencies capable of emerging in groups – small and perhaps large as well – but in the 1980s there was growing scepticism towards that possibility as well as gradual realization that 'altruism' à la Becker was not necessarily a positive force (see, for instance, Wintrobe 1983).

Building on Becker's model, B. Douglas Bernheim, Andrei Shleifer and Lawrence H. Summers (1985) included a strategic component into family transfers. The authors did not reject the possibility of sympathetic transfers from parents (or testators), but stressed above all their intention to control the beneficiaries' behaviour. In departing from Becker's model, the authors noted that the '"Ricardian equivalence" theorem' did not hold in theirs (1985, p. 1046) and that the '"rotten kid" theorem' was valid only under special circumstances (p. 1048). At least from that perspective, there was ground for reconsidering the presumed ineffectiveness of public policies.

Yet the authors preferred instead to review some macroeconomic implications of their model. When contrasted with Becker's, theirs was especially interesting because it reached the conclusion that the influence of parents over their children went further than simply dissuading opportunism within the family. While Becker's model was turned towards the absorption of the negative effects of economic and social change by the 'head' of the family, Bernheim, Schleifer and Summers, in emphasizing parents' influence on 'decisions by their children concerning education, migration and marriage' (1985, p. 1073), identified family as a factor of economic and social change. In the context of the breakdown of the traditional family unit, that conclusion could surprise, but it could also appear as the recognition that, with the loosening of family bonds, not only sympathy but also strategy was needed to prevent opportunism.

Further clarification in terms of policy implications came from Bernheim (1986) and Bernheim and Bagwell (1988), who instead of directly challenging the neutrality implications of Barro's (and Becker's) analytical framework pointed to its unsuitability to analyse the effects of public policies. In rejecting the 'Ricardian equivalence hypothesis', these authors suggested a different analytical framework in which the linkages between families, more than the 'dynastic family' à la Barro, were especially important. On the basis of these linkages, Bernheim and Bagwell (1988) established strong neutrality results, the practical implications of which they eventually dismissed on the grounds of being unrealistic. Perhaps because changes affecting family since the 1970s gained more visibility by the end of the 1980s, a number of presuppositions, characterizing Becker's and Barro's notion of family as that of a 'big happy family' behaving as if it maximized a single utility function (Bernheim and Bagwell 1988, p. 333), became gradually untenable. At the very least, the complexity of intra-family relationships seemed to call for alternative representations.

The changes in perspective can easily be realized when one considers Assar Lindbeck and Jörgen W. Weibull's (1988, p. 1165) argument about the inefficient outcomes generated by 'altruism'. In an intertemporal setting, the authors argued, gift-giving leads to social inefficiencies because the recipient can act strategically and

thus induces the donor to give more than he or she was prepared to (see also Bruce and Waldman 1990). Though reminiscent of Buchanan's 'Samaritan's Dilemma' of the mid-1970s, the argument differed in that it allowed for unselfish preferences on both sides (see also Kimball 1987). Like Buchanan's, it suggested a solution in terms of commitment à la Schelling, with the donor making a binding commitment to the level of support provided to the recipient; and, like Buchanan's, the argument included the proviso of the difficult practical enforceability of that solution. Unlike Becker's suggestion, unselfishness did not suffice to remove opportunistic tendencies in social interactions; it could even encourage them.

In the same vein, Bernheim and Stark (1988, p. 1034) saw the '"rotten kid" theorem' as rather 'special' and even identified 'a variety of circumstances in which members of a group would actually prefer to interact with less altruistic individuals, and in which the efficiency of resource allocation is inversely related to the prevailing degree of altruism' (for a perhaps more positive, though nuanced, view, see Bergstrom 1989). In addition to the criticisms levelled at Becker, that article called into question the customary distinction between family and the market in terms of behavioural assumptions. To the extent that 'altruism' tended to induce exploitability, it was suggested that 'family decisions were more properly modelled as negotiations among primarily self-interested (read: 'selfish') agents (Bernheim and Stark 1988, p. 1044). As far as society was concerned, similar conclusions apply: 'altruism' did not necessarily limit negative externalities. Worse still, unless it reached high levels, there were indications of its being a 'counterproductive social force'.

In view of the above, it may be concluded that a decade and a half after Becker and Barro had produced their results, there were serious misgivings about the generality of their application. Given that unselfishness research owed some of its impetus to the realization of the undesirable consequences of selfish behaviour in terms of the provision of public goods and considering that government intervention could be regarded as a solution to that problem, there was some irony in James Andreoni's (1990) conclusion that economic and social programmes could increase the total provision of public goods because not merely sympathetic but also selfish considerations motivated giving.

In studying privately provided public goods, Andreoni (1988) interpreted various neutrality results as many limitations of the 'pure altruism model', which he identified with the definition of the utility function of the giver as including his own consumption and the total supply of public good. Citing in passing Margolis (1982), Sugden (1984) and Bernheim et al. (1985), he called for a new approach characterized by 'non-altruistic motives for giving' (Andreoni 1988, p. 72). In subsequent works, however, Andreoni (1989, 1990) clarified his own alternative model by resorting to the warm-glow hypothesis, whereby he meant that the utility function of the giver also included his personal contribution to the public good. Combining altruism à la Tullock with altruism à la Becker, this 'impure altruism model' was said to be more consistent with empirical evidence contradicting neutrality.

Throughout the Reagan years, there were a variety of results in economics contradicting neutrality. Given the increase in the government debt over that period, it was clear that 'lesser state intervention' meant not so much strict control of federal spending as its reorientation in the context of tax reduction. From that perspective, the results obtained by Andreoni and others suggested that the existence of sympathetic transfers could not be taken as a serious justification for the ineffectiveness of national policies. Accordingly, the emphasis was shifted towards examining the power of government intervention to remedy the undesirable social consequences not only of selfish but also self-interested behaviour.

When it is remembered that Becker presented the existence of a sympathetic head as a solution to the difficulty of achieving socially desirable outcomes in various groups of otherwise selfish individuals, it is hardly surprising that the literature emphasizing the limits to 'altruism' was led to confront Becker's work on the family. In their variety, these critics did not call into question the utility maximization framework. For others,

however, that framework showed significant inadequacies when it came to explaining seemingly unselfish behaviour.

Alternative Views of Unselfishness

Just as Becker's (1974) 'A Theory of Social Interactions' epitomizes the self-interest view of unselfishness, so Sen's (1977) 'Rational Fools' represents the alternative views though the latter go beyond the well-known distinction between 'sympathy' and 'commitment'. While Sen delivered his 'Rational Fools' lecture at Oxford University in October 1976, Margaret Thatcher was already the leader of the Conservative Party and when the lecture was published in the summer of 1977 she was only a couple of years from being Prime Minister. That was a time of transition to economic liberalism. Thatcher's intention to dismantle collectivist public policies raised doubts within her own party and in society at large. The fact that Sen, a professor at the London School of Economics since 1971, proposed 'a critique of the behavioral foundations of economic theory' (the subtitle of his 1977 article) was a reminder that from the 1960s the debates on public policy in Britain had been marked by the strengthening of a vision endorsing the invisible hand of the market and economic man.

For Sen, sympathy or concern for others' welfare ('altruism' for most economists) was part of the self-interest model, whereas 'commitment' was not. He wrote:

> The former corresponds to the case in which the concern for others directly affects one's own welfare. If the knowledge of torture of others makes you sick, it is a case of sympathy; if it does not make you feel personally worse off, but you think it is wrong and you are ready to do something to stop it, it is a case of commitment. . . It can be argued that behavior based on sympathy is in an important sense egoistic, for one is oneself pleased at others' pleasure and pained at others' pain, and the pursuit of one's own utility may thus be helped by sympathetic action. It is action based on commitment rather than sympathy which would be non-egoistic in this sense. (Sen 1977, p. 326)

Perhaps because it was difficult for economists to think of an unselfish person as someone who is motivated by the welfare of others and yet benefits personally from his or her action, Sen stressed exaggeratedly both the interestedness of sympathetic agents and the indifference of committed ones.

Another aspect of Sen's approach was to link commitment to groups and then distinguish it from 'impartial concern for all', as illustrated by ethical preferences à la Harsanyi (Sen 1977, p. 336). In following that lead, Sen was echoing the earlier distinction between two sets of values, the 'economic ethic' and the 'heroic ethic', which Buchanan (1978) was now presenting under the guise of two motivational forces, 'self-interest' and 'community', the latter of which he continued to connect with group size. In the context of a changing society, which some saw as regressing economically because of the inadequate attention being given to the invisible hand mechanism, Sen felt the need to remind his readers that in addition to contributing to social harmony the economy required a degree of social cohesion and that the latter was facilitated by the individuals' sense of commitment to groups. Accordingly, economics' behavioural assumptions needed to be reconsidered so as to allow for commitment. David Collard (1978), in one of the first monographs on the subject of unselfishness, illustrated this orientation when he argued that once all self-interested motivations were allowed for, there was still room for 'a truly altruistic residual' (1978, p. 5).

By the early 1980s, it was clear that the sympathy-based view of seemingly unselfish behaviour was not the whole story. In two voluminous articles, the French economist Serge-Christophe Kolm (1981a, b) showed the complexities of 'altruism' and linked them to the prevailing schizophrenia associated with *Das Adam Smith Problem*. To some extent, Margolis's *Selfishness, Altruism, and Rationality* (1982) shifted the problem to the coexistence of two selves (or two utility functions representing an individual's self-interested preferences and his group-interested preferences, respectively) in economic man. For economists accustomed to distinguishing between economic man and moral man, Margolis's approach was disturbing. Olson, who reviewed the manuscript for Cambridge University Press, urged Margolis to reframe the argument so as to

bring it within standard economic theory (Margolis to Fontaine. 17 May 2001, personal communication), but Margolis felt that his model of individual choice was more 'consistent with the way human beings are observed to behave' (1982, p. 3).

It is unclear whether Margolis's overall approach influenced economists. Yet his distinction between 'participation altruism' – in which the economic agent gains satisfaction from giving resources away to the benefit of others – and 'goods altruism' – in which the economic agent gains satisfaction from an increase in the goods available to others – gave structure to later attempts, such as Andreoni's, to combine these two kinds of altruism.

Among the alternative views of unselfishness, the British economist Robert Sugden's (1982) deserves special mention since it proposed to reconstruct the public good theory of philanthropic behaviour, which assumed that 'the total amount of a charitable activity is an argument in the utility functions of its donors' (1982, p. 350). Having in mind the British context in which large charities exist, Sugden saw one promising option as the dropping of the utility maximization assumption and the concomitant admission that 'some individuals act on moral principles rather than on pure self-interest' (1982, p. 349). He reached the conclusion that 'the conventional argument that private philanthropy leads to the under-supply of charitable activities cannot be sustained' (1982, p. 350). In the highly charged political environment of Thatcher's first administration, such a conclusion could easily be read as another argument for lesser government intervention.

As we have seen, in the mid-1970s the ineffectiveness of economic and social policies was often justified by the existence of sympathetic transfers, but by the mid-1980s some doubted the suitability of Becker's (and Barro's) 'altruism' theories to analyse the effects of public policies. Interestingly, in a later article, Sugden (1984) explicitly dissociated his effort from 'theories of *altruism*' – by which he meant representations of behaviour in terms of concern for others. He proposed a theory of reciprocity in which, because of a Kantian rule, an individual feels obliged to make an effort (in the production of some public good) that matches others' in the group (on a more general perspective on reciprocity, see Kolm 1984). Here again, the British context was of some significance, as Sugden made clear when he mentioned the role of unpaid donors in blood procurement as an example of the supply of public goods through voluntary contributions. Sugden made the 'assumption that most people believe free riding to be morally wrong' (1984, p. 772).

The above approaches rely on groups as a relevant level of analysis between the individual and society. Recourse to ethical variables in that context makes sense as the rejection of ethics from economics has long been encouraged by its focus on impersonal relationships in the market as opposed to interactions in close-knit groups, with frequency of interactions as the main factor constitutive of sense of belongingness. More recently, however, another factor has been considered. Sen (1985), for instance, studied the influence of identification with others in the determination of a person own welfare (for an earlier attempt in that direction, see Boulding's 1962, notion of empathy in relation to groups). Sen recognized that '[o]ne of the ways in which the sense of identity can operate is through making members of a community accept certain rules of conduct as part of obligatory behavior towards others in the community' (1985, p. 349). Likewise, Herbert Simon (1992) allowed for loyalty in and identification with groups, and even accepted the working of these notions at the level of the city or nation.

In these approaches, one feels a growing uneasiness as economists move from close-knit groups, such as the family, to more informal groups, such as the country, society or humanity, in which the more obvious associations in terms of behavioural assumptions are with self-interest and not those 'perceptions of a shared humanity' which Kristen Monroe (1996) in *The Heart of Altruism* saw as central to unselfishness. There remains that in theory nothing prevents individuals from empathizing with strangers, feeling sympathy for them and behaving altruistically towards them. To date, however, this line of research has not attracted much attention.

The question may therefore be asked whether economists entertaining alternative views of

unselfishness have really been able to get over the dichotomy, to be found in the mainstream view, between the family/altruism and the market/ selfishness (see, for example, Becker 1981). Considering the slight impact of Philip Wicksteed on modern economics, it can be argued that economists have yet to digest his crucial distinction between the nature of an economic relation – the fact that the agent enters it without expressing concern for the purposes of his or her partner ('non-tuism') – and the agent's motives, which are either selfish or altruistic depending on whether the economic relation is meant to further the agent's own welfare or that of a third party (Steedman 1989; Fontaine 2000). The lack of appreciation for that distinction in modern economic theories of unselfishness and the resulting derivation of motivation (selfishness or unselfishness) from the nature of economic relation itself (impersonal or personal), explain why economists find it so unnatural to explore seemingly unselfish behaviour outside families or groups even if a number of other social scientists have shown less reluctance in that respect (see, for example, some contributions in Mansbridge 1990).

1993: Annus Mirabilis

Following attempts to investigate philanthropy in the early 1960s, unselfishness theories experienced a dramatic growth. When it is remembered that in the late 1950s economists complained about the lack of attention to love of humankind (*philanthropy*), Collard's (1992) late addition to the debate on unselfishness, 'Love is Not Enough', signalled a sea change. By early 1990, the weaknesses of research in that area could no longer be attributed to inadequate scrutiny of seemingly unselfish behaviour.

In striking contrast with the early 1960s, 1993 was a prolific year: it saw the publication of a session on the 'Economics of Altruism' in the *Papers and Proceedings* of the *American Economic Review* (Samuelson 1993; Bergstrom and Stark 1993; Simon 1993); a collection of essays, *Beyond Economic Man*, edited by Marianne Ferber and Julie Nelson (1993), which challenged the masculine foundations of economics' behavioural assumptions; and, outside economics, another collection including two essays by economists Sugden (1993) and Tyler Cowen (1993); and finally a special issue of the *Social Service Review* including interdisciplinary studies, among which was Dasgupta (1993), on the concept of 'altruism'. And to crown this achievement, Becker (1993) published a revised version of his Nobel Lecture in which he tellingly observed: 'Along with others, I have tried to pry economists away from narrow assumptions about self-interest [read: 'selfishness']. Behavior is driven by a much richer set of values and preferences' (1993, p. 385).

This list is not meant to be comprehensive, though it reflects the increasing volume of publication in this area and explains in turn the addition of an 'altruism' heading to the JEL classification system for journal articles in December 1993. Since then, research on seemingly unselfish behaviour has not slowed down, giving more room to economic experiments. There have been a reader (Zamagni 1995), several monographs and collections of essays (Stark 1995; Gérard-Varet et al. 2000) and a handbook investigating the foundations and applications of altruism research (Kolm and Mercier-Ythier 2006). If this remarkable development speaks to something it is certainly for economics' remarkable capacity to absorb and digest the most foreign subjects and notably those that present a serious challenge to its most central behavioural assumption. Whether this should be taken as a sign of strong intellectual identity is an open question.

See Also

- ▶ Altruism in Experiments
- ▶ Charitable Giving
- ▶ Economic Man
- ▶ Ethics and Economics
- ▶ Rationality, History of the Concept

Bibliography

Alchian, A.A., and W.R. Allen. 1964. *University economics*. Belmont: Wadsworth.

Andreoni, J. 1988. Privately provided public goods in a large economy: The limits of altruism. *Journal of Public Economics* 35: 57–73.

Andreoni, J. 1989. Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy* 97: 1447–1458.

Andreoni, J. 1990. Impure altruism and donations to public goods: A theory of warm-glow giving. *Economic Journal* 100: 464–477.

Arrow, K.J. 1975. Gifts and exchanges. In *Altruism, morality and economic theory*, ed. E.S. Phelps. New York: Russell Sage Foundation.

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82: 1095–1117.

Becker, G.S. 1961. Notes on an economic analysis of philanthropy. Mimeo (April). Cambridge, MA: NBER.

Becker, G.S. 1974. A theory of social interactions. *Journal of Political Economy* 82: 1063–1093.

Becker, G.S. 1976a. Altruism, egoism, and genetic fitness: Economics and sociobiology. *Journal of Economic Literature* 14: 817–826.

Becker, G.S. 1976b. *The economic approach to human behavior*. Chicago: University of Chicago Press.

Becker, G.S. 1981. Altruism in the family and selfishness in the market place. *Economica* 48: 1–15.

Becker, G.S. 1993. Nobel lecture: The economic way of looking at behavior. *Journal of Political Economy* 101: 385–409.

Becker to Tullock. 14 December 1976. Correspondence, *Journal of Economic Literature* Series, box 30, Mark Perlman Papers, Duke University, Rare Book, Manuscript, and Special Collections Library.

Bergstrom, T.C. 1989. A fresh look at the rotten kid theorem – and other household mysteries. *Journal of Political Economy* 97: 1138–1159.

Bergstrom, T.C., and O. Stark. 1993. How altruism can prevail in an evolutionary environment. *American Economic Review: Papers and Proceedings* 83: 149–155.

Bernheim, B.D. 1986. On the voluntary and involuntary provision of public goods. *American Economic Review* 76: 789–793.

Bernheim, B.D., and K. Bagwell. 1988. Is everything neutral? *Journal of Political Economy* 96: 308–338.

Bernheim, B.D., and O. Stark. 1988. Altruism within the family reconsidered: Do nice guys finish last? *American Economic Review* 78: 1034–1045.

Bernheim, B.D., A. Shleifer, and L.H. Summers. 1985. The strategic bequest motive. *Journal of Political Economy* 93: 1045–1076.

Bernstein, M.A. 2001. *A perilous progress: Economists and public purpose in twentieth-century America*. Princeton/Oxford: Princeton University Press.

Bolnick, B.R. 1975. Toward a behavioral theory of philanthropic activity. In *Altruism, morality and economic theory*, ed. E.S. Phelps. New York: Russell Sage Foundation.

Boulding, K.E. 1962. Notes on a theory of philanthropy. In *Philanthropy and public policy*, ed. F.G. Dickinson. New York: NBER.

Boulding, K.E. 1969. Economics as a moral science. *American Economic Review* 59: 1–12.

Boulding, K.E. 1978. Sociobiology or biosociology? In *Sociobiology and human nature: An interdisciplinary critique and defense*, ed. M.S. Gregory, A. Silvers, and D. Sutch. San Francisco: Jossey-Bass.

Bruce, N., and M. Waldman. 1990. The rotten-kid theorem meets the Samaritan's dilemma. *Quarterly Journal of Economics* 105: 155–165.

Buchanan, J.M. 1965. Ethical rules, expected values, and large numbers. *Ethics* 76: 1–13.

Buchanan, J.M. 1975. The Samaritan's dilemma. In *Altruism, morality and economic theory*, ed. E.S. Phelps. New York: Russell Sage Foundation.

Buchanan, J.M. 1978. Markets, states, and the extent of morals. *American Economic Review* 78: 364–368.

Collard, D. 1978. *Altruism and economy: A study in non-selfish economics*. Oxford: Martin Robertson.

Collard, D. 1992. Love is not enough. In *Thoughtful economic man: Essays on rationality, moral rules and benevolence*, ed. G. Meeks. Cambridge: Cambridge University Press.

Cowen, T. 1993. Altruism and the argument from offsetting transfers. In *Altruism*, ed. E. Frankel Paul, F.D. Miller Jr., and J. Paul. Cambridge: Cambridge University Press.

Dasgupta, P. 1993. Altruism and the allocation of resources. *Social Service Review* 67: 374–387.

Dawkins, R. 1976. *The selfish gene*, 2nd ed, 1989. Oxford/New York: Oxford University Press.

Edgeworth, F.Y. 1881. *Mathematical psychics: An essay on the application of mathematics to the moral sciences*. New York: Augustus M. Kelley. 1967.

Ferber, M.A., and J.A. Nelson. 1993. *Beyond economic man: Feminist theory and economics*. Chicago/London: University of Chicago Press.

Fontaine, P. 2000. Making use of the past: Theorists and historians on the economics of altruism. *European Journal of the History of Economic Thought* 7: 407–422.

Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*. Chicago/London: University of Chicago Press.

Frumkin, P. 1999. Private foundations as public institutions: Regulation, professionalization, and the redefinition of organized philanthropy. In *Philanthropic foundations: New scholarship, new possibilities*, ed. E. Condliffe Lagemann. Bloomington/Indianapolis: University of Indiana Press.

Gérard-Varet, L.-A., S.-C. Kolm, and J. Mercier-Ythier. 2000. *The economics of reciprocity, giving and altruism*. Basingstoke: Palgrave Macmillan.

Hall, P.D. 1999. Resolving the dilemmas of democratic governance: The historical development of trusteeship in America, 1636–1996. In *Philanthropic foundations: New scholarship, new possibilities*, ed. E. Condliffe Lagemann. Bloomington/Indianapolis: University of Indiana Press.

Hirshleifer, J. 1967. Disaster behavior: Altruism or alliance? In *Economic behaviour in adversity*. Chicago: University of Chicago Press. 1987.

Hirshleifer, J. 1977a. Economics from a biological viewpoint. *Journal of Law and Economics* 20: 1–52.

Hirshleifer, J. 1977b. Shakespeare vs. Becker on altruism: The importance of having the last word. *Journal of Economic Literature* 15: 500–502.

Hirshleifer, J. 1982. Evolutionary models in economics and law: Cooperation versus conflict strategies. *Research in Law and Economics* 4: 1–60.

Hirshleifer, J. 1985. The expanding domain of economics. *American Economic Review* 75: 53–68.

Hirshleifer to Becker. 13 December 1976. Correspondence, *Journal of Economic Literature* Series, box 31, Mark Perlman Papers, Duke University, Rare Book, Manuscript, and Special Collections Library.

Hochman, H.M., and J.D. Rodgers. 1969. Pareto optimal redistribution. *American Economic Review* 59: 542–557.

Kimball, M.S. 1987. Making sense of two-sided altruism. *Journal of Monetary Economics* 20: 301–326.

Kolm, S.-C. 1969. The optimal production of social justice. In *Public economics: An analysis of public production and consumption and their relations to the private sectors*, ed. J. Margolis and H. Guitton. London/New York: Macmillan.

Kolm, S.-C. 1981a. Efficacité et altruismes: les sophismes de Mandeville, Smith et Pareto. *Revue économique* 32: 5–31. Repr. in Kolm (1983).

Kolm, S.-C. 1981b. Altruismes et efficacités: le sophisme de Rousseau. *Information sur les sciences sociales* 20: 293–344. Repr. in Kolm (1983).

Kolm, S.-C. 1983. Altruism and efficiency. *Ethics* 94: 18–65.

Kolm, S.-C. 1984. *La Bonne Économie: La Réciprocité générale*. Paris: PUF.

Kolm, S.-C., and J. Mercier-Ythier. 2006. *Handbook of the economics of giving, altruism and reciprocity*, vol. 2 vols. Amsterdam: North-Holland.

Laffont, J.-J. 1975. Macroeconomic constraints, economic efficiency and ethics: an introduction to Kantian economics. *Economica* 42: 430–437.

Lindbeck, A., and J.W. Weibull. 1988. Altruism and time consistency: The economics of fait accompli. *Journal of Political Economy* 96: 1165–1182.

Mansbridge, J.J. 1990. *Beyond self-interest*. Chicago/London: University of Chicago Press.

Margolis, H. 1982. *Selfishness, altruism, and rationality: A theory of social choice*. Cambridge: Cambridge University Press.

Monroe, K.R. 1996. *The heart of altruism: Perceptions of a common humanity.* Princeton: Princeton University Press.

Nelson, R.R., and S.G. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, MA: Belknap.

Olson, M. 1971. *The logic of collective action: Public goods and the theory of groups*, 2nd ed. Cambridge, MA: Harvard University Press. Appendix.

Olson, M. 1982. *The rise and decline of nations: Economic growth, stagflation, and social rigidities*. New Haven/London: Yale University Press.

Phelps, E.S. 1975. *Altruism, morality, and economic theory.* New York: Russell Sage.

Phelps to Brim. 19 October 1971. Russell Sage Foundation Archives: Rockefeller Archive Center, Studies in Philanthropy, Folder 493, Box 57. New York, North Tarrytown.

Robertson, D.H. 1955. What does the economist economize? In *Economic Commentaries*. London: Staple Press. 1956.

Samuelson, P.A. 1985. Modes of thought in economics and biology. *American Economic Review: Papers and Proceedings* 75: 166–172.

Samuelson, P.A. 1993. Altruism as a problem involving group versus individual selection in economics and biology. *American Economic Review: Papers and Proceedings* 83: 143–148.

Schelling, T. 1978. Altruism, meanness, and other potentially strategic behaviors. *American Economic Review: Papers and Proceedings* 68: 229–230.

Sen, A. 1973. Behaviour and the concept of preference. *Economica* 40: 241–259.

Sen, A. 1977. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs* 6: 317–344.

Sen, A. 1985. Goals, commitment, and identity. *Journal of Law Economics & Organization* 1: 341–355.

Simon, H.A. 1990. A mechanism for social selection and successful altruism. *Science* 250: 1665–1668.

Simon, H.A. 1992. Altruism and economics. *Eastern Economic Journal* 18: 73–83.

Simon, H.A. 1993. Altruism and economics. *American Economic Review* 83: 156–161.

Stark, O. 1995. *Altruism and beyond: An economic analysis of transfers and exchanges within families and groups*. Cambridge: Cambridge University Press.

Steedman, I. 1989. Rationality, economic man and altruism. In Zamagni (1995).

Stigler, G.J. 1984. Economics – The imperial science? *Scandinavian Journal of Economics* 86: 301–313.

Sugden, R. 1982. On the economics of philanthropy. *Economic Journal* 92: 341–350.

Sugden, R. 1984. Reciprocity: The supply of public goods through voluntary contributions. *Economic Journal* 94: 772–787.

Sugden, R. 1993. Thinking as a team: Towards an explanation of nonselfish behavior. In *Altruism*, ed. E. Frankel Paul, F.D. Miller Jr., and J. Paul. Cambridge: Cambridge University Press.

Thurow, L.C. 1971. The income distribution as a pure public good. *Quarterly Journal of Economics* 85: 327–336.

Tullock, G. 1977. Economics and sociobiology: A comment. *Journal of Economic Literature* 15: 502–506.

Tullock, G. 1978. Altruism, malice and public goods. *Journal of Social and Biological Structures* 1: 3–9.

Tullock, G. 1979. Sociobiology and economics. *Atlantic Economic Journal* 7: 1–10.

Tullock, G. 1990. The economics of (very) primitive societies. *Journal of Social and Biological Structures* 13: 151–162.

Vickrey, W.S. 1962. One economist's view of philanthropy. In *Philanthropy and public policy*, ed. F.G. Dickinson. New York: NBER.

Wilson, E.O. 1975. *Sociobiology*, Abridged edn. Cambridge, MA: Belknap, 1980.
Wintrobe, R. 1983. Taxing altruism. *Economic Inquiry* 21: 255–270.
Zamagni, S. 1995. *The economics of altruism*. Aldershot/Brookfield: Edward Elgar.

A

Ambiguity and Ambiguity Aversion

Marciano Siniscalchi

Abstract

Experimental evidence strongly suggests that subjects facing a decision under uncertainty often find it difficult to assess the relative likelihood of certain events; decision theorists deem such events 'ambiguous'. Furthermore, subjects generally dislike options (acts) whose final outcome depends upon the realization of such ambiguous events; that is, they are 'ambiguity-averse'. This article surveys the main decision-theoretic models developed since the mid-1980s to accommodate ambiguity and ambiguity aversion, including Choquet-expected utility (Schmeidler. Econometrica 57: 571–87, 1989) and maxmin expected utility (Gilboa and Schmeidler. J Math Econ 18: 141–53, 1989). More recent developments in the theory of ambiguity are also briefly summarized.

Keywords

Ambiguity; Ambiguity aversion; Bernoulli utility function; Choice under uncertainty; Choquet-expected utility; Decision theory; Decision weights; Ellsberg paradox; Incomplete preferences; Maximum likelihood; Maxmin expected utility; Multiple priors; Preference reversals; Probability; Savage, L; Second-order probabilities; Subjective expected utility; Sure-thing principle; von Neumann–Morgenstern utility function

JEL Classifications

D1; D8

Consider the following choice problem, known as 'Ellsberg's three-colour urn example', or simply the 'Ellsberg paradox' (Ellsberg 1961). An urn contains 30 red balls, and 60 green and blue balls, in unspecified proportions; subjects are asked to compare (*a*) a bet on a red draw with a bet on a green draw, and (*b*) a bet on a red or blue draw with a bet on a green or blue draw. If the subject wins a bet, she receives ten dollars; otherwise, she receives zero dollars. To model this situation as a problem of choice under uncertainty, let the state space be $\{s_r, s_g, s_b\}$, in obvious notation, and consider the bets in Fig. 1.

The modal preferences in this example are $f_r \succ f_g$ and $f_{rb} \prec f_{gb}$, where '$\succ$' denotes strict preference. (Ellsberg did not conduct actual experiments, but similar patterns of behaviour have been reported in subsequent experimental studies; see Camerer and Weber 1992, for an exhaustive survey.) A common rationalization runs as follows: betting on red is 'safer' than betting on green, because the urn may actually zero green balls; on the other hand, betting on green or blue is 'safer' than betting on red or blue, because the urn may contain zero blue balls. Equivalently, when one evaluates f_r and f_{gb}, the fact that the relative likelihood of green as against blue balls is unspecified is irrelevant; on the other hand, this consideration looms large when one evaluates the acts f_g and f_{rb}.

While these preferences seem plausible, they are inconsistent with subjective expected utility maximization (SEU). Indeed, they are inconsistent with the weaker assumption that the decision-maker's (DM) qualitative beliefs, as revealed by her betting behaviour, can be numerically

	S_r	S_g	S_b
f_r	10	0	0
f_g	0	10	0
f_{rb}	10	0	10
f_{gb}	0	10	10

Ambiguity and Ambiguity Aversion, Fig. 1 Ellsberg's three-colour urn

represented by a probability measure. Note that $f_r \succ f_g$ indicates that r is deemed strictly more likely than g, so any probability P that represents the individual's likelihood ordering of events must satisfy $P(\{r\}) > P(\{g\})$; on the other hand, $f_{rb} \prec f_{gb}$ indicates that $\{r, b\}$ is strictly less likely than $\{g, b\}$, which would require $P(\{r\}) + P(\{b\}) = P(\{r, b\}) < P(\{g, b\}) = P(\{g\}) + P(\{b\})$, hence $P(\{r\}) < P(\{g\})$.

The key to Ellsberg's example is the fact that the composition of the urn is incompletely specified; in particular, the relative likelihood of a green as against a blue draw is 'ambiguous'. More generally, in the words of Daniel Ellsberg, *ambiguity* is

> a quality depending on the amount, type, reliability and 'unanimity' of information, and giving rise to one's 'degree of confidence' in an estimate of relative likelihoods. (1961, p. 657).

To borrow Ellsberg's terminology, the modal preferences $f_r \succ f_g$ and $f_{rb} \prec f_{gb}$ indicate that the DM would rather have the ultimate outcome of her choices (that is, whether she receives 10 or 0) depend upon events about whose relative likelihood she is more confident. In other words, these preferences denote *ambiguity aversion*.

Since the mid-1980s, several decision models that can accommodate ambiguity and ambiguity aversion (or appeal) have been axiomatized; other contributions have addressed the behavioural manifestations and implications of ambiguity, as well as updating and dynamic choice. Furthermore, there is an ever-growing collection of applications to contract theory, auctions, finance, macroeconomics, political economy, insurance and other areas of economic inquiry.

The following section reviews two of the most influential models of ambiguity-sensitive preferences in a static setting, while the succeeding section briefly discusses additional models, updating, and dynamic choice.

'Classical' Models of Ambiguity-Sensitive Preferences

Preliminaries

Fix a finite or infinite state space S and an algebra Σ of its subsets. A *probability charge* is set function $P : \Sigma \to [0, 1]$ that satisfies $P(S) = 1$ and $P(E \bigcup F) = P(E) + P(F)$ for all $E, F \in \Sigma$ with $E \cap F = \emptyset$; that is, P is normalized and *finitely* additive. The set of probability charges on (S, Σ) is denoted $\Delta(S, \Sigma)$.

The decision models discussed in this section were first axiomatized in the framework introduced by Anscombe and Aumann (1963); it is convenient to adopt the same set-up here. (Alternative axiomatizations that do not rely on lotteries have also been obtained: see, for example, Gilboa 1987; Chew and Karni 1994; Casadesus-Masanell et al. 2000; Ghirardato et al. 2003). Fix a set of prizes X, and let $\Delta(X)$ be the collection of all lotteries (probability distributions) on X with finite support. An act is a Σ-measurable map $f : S \to \Delta(X)$. The set $\Delta(X)$ is closed under mixtures, that is, convex combinations; mixtures of acts are then defined pointwise, so that the set $\mathcal{F}$ of all acts is also closed under mixtures (that is, for every α A $\in$ [0, 1] and every pair of acts f, g, $\alpha f + (1 - \alpha)g$ is the act that yields the lottery $\alpha f(s) + (1 - \alpha)g(s)$ in state $s \in S$).

A preference is a binary relation $\succeq$ on $\mathcal{F}$; its symmetric and asymmetric parts are denoted by $\sim$ and $\succ$ respectively. It is customary to identify every lottery $p \in \Delta(X)$ with the constant act that yields p in every state.

A (von Neumann–Morgenstern, or Bernoulli) utility function is a map $u : \Delta(X) \to \mathbb{R}$ that satisfies $u(\alpha p + (1 - \alpha)q) = \alpha u(p) + (1 - \alpha)u(q)$ for all $\alpha \in [0, 1]$ and $p, q \in \Delta(X)$. All axiomatizations discussed below ensure that preferences over lotteries can be represented by a utility function.

A function $a : S \to \mathbb{R}$ is simple if its range is finite; write $a = (a_1, E_1, \ldots, a_n, E_n)$, where $a_1, \ldots, a_N \in \mathbb{R}$ and $E_1, \ldots, E_N$ is a partition of S, to indicate that, for all $n = 1, \ldots, N$, $a(s) = a_n$ for all $s \in E_n$. An act is simple if its range can be partitioned into finitely many indifference classes. The set of simple Σ-measurable acts is denoted by $\mathcal{F}_0$.

Virtually all substantive decision-theoretic issues can be analysed by restricting attention to preferences over $\mathcal{F}_0$; the reader is urged to consult the references cited for a discussion of preferences over non-simple acts.

Capacities and Choquet-Expected Utility

The modal preferences in the three-colour urn example are inconsistent with a probabilistic representation of beliefs essentially because probabilities are finitely additive. Specifically, if the probability charge P represents the individual's qualitative beliefs, $f_{rb} \prec f_{gb}$ requires that $P(\{r, b\}) < P(\{g, b\})$; since P is additive, this implies $P(\{r\}) < P(\{g\})$. However, $f_r \prec f_g$ implies the reverse inequality. Thus, formally, the Ellsberg paradox can be 'resolved' if a weaker, non-additive representation of the individual's qualitative beliefs is allowed. This approach is pursued in Schmeidler (1986, 1989).

A *capacity* is a set function $v : \Sigma \to [0, 1]$ such that $v(S) = 1$ and $v(A) \leq v(B)$ for all events A, $B \in \Sigma$ such that $A \subseteq B$. Thus, a capacity is not required to be additive, although it must satisfy a monotonicity property that has a natural interpretation in terms of qualitative beliefs: 'larger' events are 'more likely'.

To define expectation with respect to capacities, a suitable notion of integration is required. Consider a simple function $a = (a_1, E_1, \ldots, a_N, E_N)$, with $a_1 > a_2 > \ldots > a_N$.The *Choquet integral* of a with respect to a capacity v (Choquet 1953) is the quantity

$$\int a\,dP = \sum_{n=1}^{N-1} (a_n - a_{n+1}) v\left(\bigcup_{m=1}^{n} E_m\right) + a_N. \quad (1)$$

With the convention that $\cup_{m=1}^{0} E_m = \varnothing$, Eq. (1) can be rewritten as follows:

$$\int a\,dP = \sum_{n=1}^{N} a_n \left[v\left(\bigcup_{m=1}^{n} E_m\right) - v\left(\bigcup_{m=1}^{n-1} E_m\right)\right]. \quad (2)$$

Thus, Choquet integration performs a 'weighted average' of the values $a_1, \ldots, a_N$, with non-negative weights $v(E_1)$, $v(E_1 \cup E_2) - v(E_1)$, $\ldots$, $1 - v(E_1 \cup \ldots \cup E_{N-1})$ that add up to one. If v is additive, Eq. (1) reduces to $\int a\,dP = \sum_{n=1}^{N} a_n v(E_n)$. However, in general, the ordering of the values $a_1, \ldots, a_N$ affects the decision weights: for instance, suppose $a = (\alpha, E, \beta, S/E)$, with $\alpha \neq \beta$: then $\int a\,dv$ equals $\alpha v(E) + \beta[1 - v(E)]$ if $\alpha > \beta$, and $\beta v(S/E) + \alpha[1 - v(S/E)]$ if $\beta > \alpha$. These expressions are different unless $v(E) + v(S/E) = 1$.

A preference admits a *Choquet-expected utility* (CEU) representation if there exists a utility function u and a capacity v such that, for all simple acts $f, g \in F_0$, $f \succeq g$ if and only if $\int u(f(s))\,dv \geq \int u(g(s))\,dv$, where the integrals are as in Eq. (1).

Preferences in the Ellsberg paradox are consistent with CEU. Let u satisfy $u(10) > u(0)$, and observe that $f_r \succ f_g$ requires $v(\{r\}) > v(\{g\})$, whereas $f_{rb} \prec f_{gb}$ implies that $v(\{r, b\}) < v(\{g, b\})$; since v is not required to be additive, these inequalities can be mutually consistent: for instance, let

$$\begin{aligned} v(\{r\}) = v(\{r, b\}) = v(\{r, g\}) \\ = \frac{1}{3}, v(\{b\}) = v(\{g\}) \\ = 0, \text{and } v(\{b, g\}) = \frac{2}{3}. \end{aligned} \quad (3)$$

Recall that the key axiom in the Anscombe–Aumann axiomatization of SEU is *Independence*: for all triples of (simple) acts f, g, h, and all $\alpha \in (0, 1)$, $f \succ g$ implies $\alpha f + (1 - \alpha)h \succ \alpha g + (1 - \alpha)h$. Schmeidler (1989) shows that CEU preferences are instead characterized by a weaker independence property. Say that two acts f and g are *comonotonic* if there is no pair of states s, s' such that $f(s) \succ f(s')$ and $g(s) \prec g(s')$; the key axiom in Schmeidler's characterization of CEU preferences, *Comonotonic Independence*, requires that $f \succ g \Rightarrow \alpha f + (1 - \alpha)h \succ \alpha g + (1\,\alpha)h$ only if f, g, h are pairwise comonotonic.

To illustrate the rationale behind this weakening of Independence, consider the acts f_r and f_g in the Ellsberg paradox, and define a third act f_b by $f_b(r) = f_b(g) = 0$ and $f_b(b) = 10$. For the CEU preferences defined above, $f_r \succ f_g$, but $\frac{1}{2}f_r + \frac{1}{2}f_b \prec \frac{1}{2}f_g + \frac{1}{2}f_b$. This is consistent with the notion that the DM dislikes ambiguity, and hence would rather have the ultimate outcome of her choices depend upon events about whose relative likelihood she is more confident; in particular, notice that the mixture $\frac{1}{2}f_g + \frac{1}{2}f_b$ yields the same

outcome in states g and b, so the DM need not worry about her lack of confidence in her assessment of their relative likelihood.

This example also suggests that mixtures of non-comonotonic acts can be appealing for an individual who might informally be described as 'ambiguity-averse'. As was just noted, mixtures of f_g and f_b can reduce or eliminate the dependence of the final outcome upon the realization of g rather than b, and hence provide a *hedge against ambiguity*. The DM under consideration finds this appealing: $\frac{1}{2}f_g + \frac{1}{2}f_b \succ f_g \sim f_b$.

Schmeidler (1989) suggests that this 'preference for mixtures' may be taken as a behavioural definition of *ambiguity aversion*. Formally, say that an individual is ambiguity-averse if, for all $f, g \in \mathscr{F}_0, f \succeq g$ implies $\alpha f + (1 - \alpha)g \succeq g$. Schmeidler then shows that a CEU individual is ambiguity-averse if and only if the capacity representing her preferences is *convex*: that is, for all events E, $F \in \Sigma$, $\nu(E \cup F) + \nu(E \cap F) \geq \nu(E) + \nu(F)$. For instance, the capacity in Eq. (3) is convex.

Multiple Priors and Maxmin Expected Utility

Gilboa and Schmeidler (1989, p. 142) propose an alternative rationalization of the preferences $f_r \succ f_g$ and $f_{rb} \prec f_{gb}$ in the Ellsberg paradox:

> One conceivable explanation of this phenomenon which we adopt here is as follows:. . .the subject has too little information to form a prior. Hence (s)he considers a *set* of priors as possible. Being [ambiguity] averse, s(he) takes into account the *minimal* expected utility (over all priors in the set) while evaluating a bet. For an analysis of this interpretation of multiple priors, see Siniscalchi (2006).

Formally, preferences admit a *maxmin expected utility* (MEU) decision rule if, given a utility function u and a weak* closed, convex set C of probability charges on S, for all $f, g \in \mathscr{F}_0, f \succeq g$ if and only if

$$\min_{P \in C} \int u(f)\, dP \geq \min_{P \in C} \int u(g)\, dP,$$

where integration has the usual meaning. For instance, the modal rankings in the Ellsberg paradox are consistent with MEU, with $u(10) > u(0)$ and

$$C = \left\{ P \in \Delta(S, \Sigma) : P(\{r\}) = \frac{1}{3} \right\} \qquad (4)$$

(other choices of C are possible).

Gilboa and Schmeidler's axiomatization of the MEU decision rule features two key axioms: *C-Independence* and Ambiguity Aversion. The latter was stated in the previous subsection; C-Independence requires that, for all acts $f, g \in \mathscr{F}_0$ and all *constant* acts, or lotteries, $p \in \Delta(X)$, $f \succeq g$ if and only if $\alpha f + (1 - \alpha)p \succeq \alpha g + (1 - \alpha)p$. Thus, relative to the full Independence axiom, preference reversals are ruled out only for mixtures with constant acts.

Intuitively, mixing an act with a constant does not provide any hedging opportunities; rather, such mixtures change only the 'scale and location' of an act's utility profile. Thus, the requirement formalized by C-Independence is consistent with the discussion in the preceding subsection; indeed, CEU preferences satisfy C-Independence. On the other hand, C-Independence allows for violations of Comonotonic Independence (see Klibanoff 2001, for an example and further discussion).

Ambiguity-averse CEU preferences satisfy both C-Independence and Ambiguity Aversion (in addition to other structural axioms); thus, they are MEU preferences. Schmeidler (1989) shows that, in particular, the convex capacity ν representing an ambiguity-averse CEU preference is the *core* of the set C of priors in the MEU representation of the same preferences: that is, $C = \{P \in \Delta(S, \Sigma) : \forall E \in \Sigma, P(E) \leq \nu(E)\}$. For instance, the capacity ν in Eq. (3) is the core of the set C in Eq. (4).

Other Models, Updating, and Dynamic Choice

A generalization of the MEU model, related to Hurwicz's α-maxmin criterion (cf. Luce and Raiffa 1957, p. 304), sometimes appears in applications; given a utility function u, a weak*-closed, convex set C of priors, and a number $\alpha \in [0, 1]$, $f \succeq g$ if and only if

$$\alpha \min_{P \in C} \int u(f)\, dP + (1-\alpha) \max_{P \in C} \int u(f)\, dP$$
$$\geq \alpha \min_{P \in C} \int u(g)\, dP + (1-\alpha) \max_{P \in C} \int u(g)\, dP;$$

thus, MEU corresponds to the case $\alpha = 1$. An axiomatization and further discussion can be found in Ghirardato et al. (2004).

Truman Bewley (2002) proposes an alternative approach to ambiguity. In both the CEU and MEU models, the DM responds to ambiguity by essentially evaluating different acts using different 'decision weights'. Bewley suggests that, alternatively, the DM may simply be unable to rank certain acts in the presence of ambiguity; in other words, preferences may be *incomplete*. He axiomatizes the following partial decision rule: for a given utility function u and weak* closed, convex set C of priors, $f \succeq g$ if and only if

$$\forall P \in C, \int u(f)\, dP \geq \int u(g)\, dP.$$

For instance, in Ellsberg's three-colour-urn example, if the set C is chosen as above, the DM is unable to rank the acts f_r and f_g, as well as the acts f_{rb} and f_{gb}. Notice that preferences satisfy the full Independence axiom in Bewley's model: ambiguity manifests itself solely through incompleteness.

Ambiguity can also be modelled by introducing *second-order probabilities*. For instance, Klibanoff et al. (2005) axiomatize the following decision rule:

$$\forall f, g \in F_0, f \succeq g \Leftrightarrow \int_{\Delta(S)} \varphi\left(\int_S u(f)\, dP\right) d\mu \geq \int_{\Delta(S)} \varphi\left(\int_S u(g)\, dP\right) d\mu,$$

where μ is a probability measure over the set $\Delta(S)$ of probability charges on the finite state space S, and φ is a 'second-order utility function'. A notion of ambiguity aversion is characterized by concavity of φ. See also Ergin and Gul (2004).

Recent contributions aim at characterizing ambiguity without restricting attention to specific decision models, and without relying on functional-form considerations. Epstein and Zhang (2001) propose a definition of 'unambiguous event' that is based solely on preferences. Under suitable structural axioms, preferences over acts that are measurable with respect to such 'subjectively unambiguous' events are *probabilistically sophisticated* in the sense of Machina and Schmeidler (1992); this indicates that the proposed behavioural definition characterizes absence of ambiguity. See also Epstein (1999) for a related assessment of Schmeidler's definition of ambiguity aversion.

Ghirardato et al. (2004) note that, in models such as CEU and MEU, ambiguity manifests itself via violations of the Anscombe–Aumann Independence axiom. Thus, they propose to deem an act f 'unambiguously preferred' to an act g if $\alpha f + (1-\alpha)h \succeq \alpha g + (1-\alpha)h$ for all $\alpha \in (0, 1)$ and all $h \in F_0$. They show that unambiguous preference admits a Bewley-style representation, characterized by a set C of priors which is a singleton if and only if the original preference is SEU. In light of this result, they suggest that the DM perceives ambiguity whenever C is not a singleton. See also Ghirardato and Marinacci (2002).

To highlight the differences between these definitions, consider a probabilistically sophisticated, non-SEU preference. According to the Epstein–Zhang definition, all events are subjectively unambiguous, whereas the Ghirardato–Maccheroni–Marinacci approach concludes that some ambiguity is perceived.

The modal preferences in the Ellsberg paradox constitute a violation of the *surething principle*, which is arguably the centrepiece of Leonard Savage's (1954) axiomatization of SEU; indeed, this was a main focus of Ellsberg's seminal article. However, the sure-thing principle also plays a key role in ensuring that conditional preferences are well-defined and 'dynamically consistent'; finally, it provides a foundation for Bayesian updating. Thus, since ambiguity leads to violations of the sure-thing principle, defining updating and ensuring a suitable form of dynamic consistency for MEU, CEU and similar decision models presents some challenges.

Gilboa and Schmeidler (1993) axiomatize Dempster–Shafer updating of capacities

(cf. Dempster 1968; Shafer 1976) and 'maximum-likelihood updating' of multiple priors for ambiguity-averse CEU preferences. Prior-by-prior updating for MEU preferences is axiomatized in Jaffray (1994).

All these updating rules may lead to 'dynamic inconsistencies', that is, preference reversals: the ranking of two acts may be different before and after learning than a (typically ambiguous) event has occurred. Epstein and Schneider (2001) instead axiomatize a model of recursive MEU preferences by explicitly imposing dynamic consistency with respect to a pre-specified filtration. The recursive formulation is especially convenient in applications; on the other hand, dynamic consistency imposes some restrictions on the set of MEU priors: see Epstein and Schneider (2001) for further discussion. Wang (2003) provides related results. Dynamic choice under ambiguity is currently an area of active research.

See Also

- Decision Theory in Econometrics
- Expected Utility Hypothesis
- Measure Theory
- Non-Expected Utility Theory
- Risk Aversion
- Savage's Subjective Expected Utility Model
- Uncertainty

Bibliography

Aliprantis, C., and K. Border. 1994. *Infinite dimensional analysis*. Berlin: Springer.

Anscombe, F., and R. Aumann. 1963. A definition of subjective probability. *Annals of Mathematical Statistics* 34: 199–205.

Bewley, T. 2002. Knightian decision theory: Part I. *Decisions in Economics and Finance* 25(2): 79–110.

Camerer, C., and Martin Weber. 1992. Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty* 5: 325–370.

Casadeus-Masanell, R., P. Klibanoff, and E. Ozdenoren. 2000. Maxmin expected utility over savage acts with a set of priors. *Journal of Economic Theory* 92: 33–65.

Chew, H., and E. Karni. 1994. Choquet expected utility with a finite state space: Commutativity and act-dependence. *Journal of Economic Theory* 62: 469–479.

Choquet, G. 1953. Theory of capacities. *Annales de l'Institut Fourier (Grenoble)* 5: 131–295.

Dempster, A. 1968. A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Series B* 30: 205–247.

Ellsberg, D. 1961. Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics* 75: 643–669.

Epstein, L. 1999. A definition of uncertainty aversion. *Review of Economic Studies* 66: 579–608.

Epstein, L., and M. Schneider. 2001. Recursive multiple-priors. *Journal of Economic Theory* 113: 1–31.

Epstein, L., and J. Zhang. 2001. Subjective probabilities on subjectively unambiguous events. *Econometrica* 69: 265–306.

Ergin, H., and F. Gul. 2004. A subjective theory of compound lotteries. Mimeo. *Econometric society 2004 North American summer meetings*, No. 152.

Ghirardato, P., and M. Marinacci. 2002. Ambiguity made precise: A comparative foundation. *Journal of Economic Theory* 102: 251–289.

Ghirardato, P., F. Maccheroni, M. Marinacci, and M. Siniscalchi. 2003. A subjective spin on roulette wheels. *Econometrica* 71: 1897–1908.

Ghirardato, P., F. Maccheroni, and M. Marinacci. 2004. Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory* 118: 133–173.

Gilboa, I. 1987. Expected utility with purely subjective non-additive probabilities. *Journal of Mathematical Economics* 16: 65–88.

Gilboa, I., and D. Schmeidler. 1989. Maxmin expected utility with a non-unique prior. *Journal of Mathematical Economics* 18: 141–153.

Gilboa, I., and D. Schmeidler. 1993. Updating ambiguous beliefs. *Journal of Economic Theory* 59: 33–49.

Jaffray, J.-Y. 1994. Dynamic decision making with belief functions. In *Advances in the Dempster–Shafer theory of evidence*, ed. R. Yager, J. Kacprzyk, and M. Fedrizzi. New York: Wiley.

Klibanoff, P. 2001. Characterizing uncertainty aversion through preference for mixtures. *Social Choice and Welfare* 18: 289–301.

Klibanoff, P., M. Marinacci, and S. Mukerji. 2005. A smooth model of decision making under ambiguity. *Econometrica* 73: 1849–1892.

Luce, R., and H. Raiffa. 1957. *Games and decisions*. New York: Wiley.

Machina, M., and D. Schmeidler. 1992. A more robust definition of subjective probability. *Econometrica* 60: 745–780.

Savage, L. 1954. *The foundations of statistics*. New York: Wiley.

Schmeidler, D. 1986. Integral representation without additivity. *Proceedings of the American Mathematical Society* 97: 255–261.

Schmeidler, D. 1989. Subjective probability and expected utility without additivity. *Econometrica* 57: 571–587.

Shafer, G. 1976. *A mathematical theory of evidence*. Princeton: Princeton University Press.
Siniscalchi, M. 2006. A behavioral characterization of plausible priors. *Journal of Economic Theory* 128(1): 91–135.
Wang, G. 2003. Conditional preferences and updating. *Journal of Economic Theory* 108: 286–321.

American Economic Association

Michael A. Bernstein

Abstract

Following its foundation, the American Economic Association (AEA) cultivated a unique professional visibility, struggling to establish its professional credentials and to demonstrate the usefulness of economists' ostensible skills. While celebrating the virtues of 'free markets', the AEA was itself shaped by government and collective action. In the latter half of the 20th century, the AEA promoted a 'New Economics' focused on macroeconomic intervention and regulation. However, these developments fostered a new generation of specialists with different views of public purpose, the appropriate role of government, and how professional economists could participate in the formulation and implementation of public policy.

Keywords

American Economic Association; American Finance Association; American Home Economics Association; Business cycles; Economics, definition of; Home economics; Royal Economic Society; Union for Radical Political Economics

JEL Classifications

B1

The American Economic Association (AEA) was inaugurated by a miscellaneous group of scholars, university administrators and public figures, in September 1885, in the early stages of a sustained expansion in American academic life. Its original objectives of encouraging research, publications on economic subjects, and perfect freedom in economic discussions have been consistently maintained, sometimes not without difficulty given the disagreements among its members, and the persistent tension between the desire for scientific objectivity and non-partisanship on the one hand and the urge to make an impact on public policy on the other. This problem was especially acute during the AEA's early years, when economic questions were at the forefront of public discussion. A number of prominent American economists were then under attack, and some were dismissed from or forced out of their university posts because of their opinions. However, under its first President, F.A. Walker, an internationally known figure who served for the first seven years, the AEA gradually lost some of its initial reformist tone and concentrated increasingly on more strictly scholarly issues. Unlike the British Royal Economic Society, which has frequently had a non-professional president, the AEA has invariably been dominated by academic economists, although in recent decades prominent government professional economists have occasionally held the office – for example, Alice Rivlin, the first woman President, in 1985.

Early Challenges and Strategies

While the AEA's contributions to economic knowledge through its periodicals – the *American Economic Review* (from 1911), the *Journal of Economic Literature* (from 1963), and the *Journal of Economic Perspectives* (from 1987) – and in various other ways are undeniable, its services to the profession have perhaps been unnecessarily restricted because of the heterogeneity of its constituency, which has always included a substantial proportion of non-academic members, and its commitment to nonpartisanship. Thus, for example, the AEA's reactions to the conflicts and tensions in American society have been distinctly more cautious than those of some other learned societies, both within and outside the social

sciences, with respect to academic freedom issues. However, in both world wars the AEA played a notable and constructive part by organizing professional expertise for government service, and by conducting open debates and issuing publications on the economic problems of war and peace. The Association has also since 1945 occupied a leading role in the internationalization of the economic profession. It has always been an 'open' society, with no significant membership restrictions, partly because of the objections to control by a limited elite or coterie. Consequently it has only occasionally had any direct influence on doctrinal developments in the field. Nevertheless, there have been periodic protests about the organization's unrepresentativeness and oligarchic management, a state of affairs reflecting the size, diversity, and geographical dispersion of its membership, which now stands at a little over 22,000 (including subscribers).

Under its charter of incorporation, the AEA committed itself to 'the encouragement of economic research, especially the historical study of the actual conditions of industrial life' as well as to 'the encouragement of perfect freedom of economic discussion'. In particular, 'the Association as such [took] no partisan attitude, nor commit [ed] its members to any position on practical economic questions'. While the formal organization was thus made distinct from the individual activities and convictions of its members, nevertheless the stresses and strains attendant upon the struggles over its initial establishment were, in its earliest years, never far from the surface. These anxieties in turn framed the process by which major decisions were ultimately made concerning AEA membership criteria, annual meetings, publications, and operational procedures; what is more, they made the Association's leadership particularly eager to seize upon whatever opportunities and circumstances within the public arena might enhance the prestige and sway of their field.

From its earliest days, the AEA faced certain difficulties associated with maintaining the separation between professional image and individual values. One of these involved continuing struggles over academic freedom issues, involving economists at certain educational institutions across the nation. The most celebrated of these, although by no means the only ones, were the cases of Richard Ely at the University of Wisconsin, Edward Bemis at the University of Chicago, and Edward Ross at Stanford. All three scholars had been accused in the 1890s, in different contexts and in various ways, of poisoning the minds of their students with ideas and beliefs inimical to corporate interests and private wealth. Two of them, Ely and Ross, managed to bring their careers back from the brink of the abyss; Bemis was not as fortunate and, in the end, was condemned to oblivion. Whether in success or failure, however, the defence of colleagues placed in jeopardy for their political convictions and beliefs relied more on the *individual* support of powerful champions within the profession rather than on the *collective* imprimatur of the AEA.

Fretting over the size of their professional society was, for the early AEA leadership, one thing; firmly articulating the Association's *raison d'être* was something else. Declarations of purpose, no matter how frequently or even stridently made, served only to a point. It was in actual practice, and in the decisions that animated it, that the professional community of the AEA truly explained and revealed itself. No amount of enforcement of particular boundaries of expertise could substitute for the rigorous refinement of colleagues that would result from the inculcation of specific ways of doing the community's business. Whether self-consciously or not, Association members and officials were, from the earliest years of the 20th century, concerned to frame the interests, activities, and procedures of their group in ways that would, more powerfully and vividly than any set of membership standards might, decisively create and preserve the profession that it was their goal to foster.

Creating a professional journal was also quite challenging. With no debate among AEA secretariat colleagues, Davis Dewey, the founding editor of the *American Economic Review*, rejected a suggestion from the Theodora B. Cunningham in 1916 that the journal include 'a Women's Department of household economics'. Dewey's decision in this regard was thoroughly consistent with not one but two strategies of professionalization in

early-20th-century America. On the one hand, it furthered the conscious effort of AEA founders to secure a distinctive place for economics as a scientifically grounded enterprise that avoided the lesser prestige of feminized occupations like 'home economics'. On the other, it actually dovetailed with efforts dating from 1900 to constitute home economics as a separate discipline in its own right. Women professionals eager to find in the home economics field the same authority and influence that their male counterparts struggled for in an array of other disciplines had worked assiduously to establish collegiate degree programmes, journals, and a national association – the American Home Economics Association (AHEA). Their very success made the 'defeminization' of economics, at the hands of professional communities like the AEA, rather easy.

In fact, the question of publication standards threatened to destabilize the general consensus about the desirability of creating the *American Economic Review* in the first place. Argument over the implementation of standards not only raised questions of intellectual freedom and openness but also drew attention back to the general and often delicate matter of the journal's purpose. Not simply value as to method and technique, but significance and appropriateness as to subject figured prominently in the deliberations of the AEA Executive Council regarding the new journal and the Association's annual meetings. These discussions continued for years and ultimately decades to come. They were, in fact, often intertwined, touching upon related concerns about professional status and prestige, scientific conduct and codes, and the boundaries (topical and methodological) of economics itself. Stoutly defining what economics was involved being clear-minded about what it was not. Prominent AEA members, at the very moment they were wrestling with the nature of a new publication for the Association, vigorously protested to President Seligman that sociologists be kept at bay from the annual meeting and even the quarterly itself. 'We have heard [the sociologists] so many times', Henry Carter Adams wrote Seligman in the spring of 1902, 'that we know absolutely what each one of the[m] will say upon any subject'. When gathered in an annual convention, Thomas Carver argued, 'Economists would prefer to stick to the subject of Economists. [One] should especially doubt whether the members of [the] association would easily find a common ground of discussion with Miss [Jane] Addams or Mr. Felix Adler, admirable as these persons are and valuable as their work is. [One] should be afraid that there would be difficulty in trying to think in the same language.' The same, Carver believed, was true for the *Review.* He doubted very much if 'it would be wise to include much sociology, except such as has a distinctly economic coloring'. (All quotations of AEA minutes and correspondence are from the AEA Archives, Northwestern University Library, Box 8.)

Enforcing disciplinary boundaries, in both publication strategies and convention planning, also involved making precise decisions about the relationship between scholarly research and contemporary policy debate. With apparently little discussion or debate, the AEA Executive Committee formally chose in 1915 to exclude from the pages of the *American Economic Review* a 'department of current economic events'. Even if contemporary policy concerns found their way into the submissions to the Association's quarterly, the editors were determined 'that current economic questions . . . be treated by scholarly men and not left to the sensational magazine writer'. In some respects this was a curious position for the leadership to assume given the additional concern that the work of economists be made visible and influential in the world of public affairs. The notion that the *Review* should be 'a craftsman's tool' had, after all, animated a great deal of the effort of the editorial office from the earliest days. Maintaining a dispassionate, scholarly tone while encouraging a wide and even diverse readership was neither a simple nor an obvious task. Editor Davis Dewey put it well to the distinguished English theorist Francis Edgeworth in January 1911 when he wrote, 'We are trying to appeal to a somewhat varied membership who are interested in current questions. We do not, however, wish to be popular in a commonplace way, but shall endeavor to have our articles prepared by men of scholarly standards.'

The problem of attracting 'a somewhat varied membership' while adhering to 'scholarly standards' that would guard against being 'popular in a commonplace way' was truly vexing.

The Impact of National Mobilizations and Emergencies

The coming of the Great War stimulated the professionalization of AEA ranks. In the spring of 1914, the AEA secretariat fashioned a special opportunity to bring the potential benefits of professional economics expertise to the attention of federal officials. Not surprisingly, it involved concerns with the ways in which the United States Department of Agriculture (DOA) calculated and reported statistical data on the performance of the nation's farms. Cornell University Professor Allyn Young contacted the secretary of agriculture, David F. Houston, to express the fear of the AEA leadership that 'much of the statistical work . . . issu [ed] from government offices [wa]s of disgracefully poor quality'. He noted that the failures of the DOA in this regard were by no means unique. Clearly, 'many of the activities of [federal] government bureaus furnish[ed] statistical by-products that [c]ould be of the greatest usefulness'. There was a clear need, in Young's opinion, that these data be 'properly tabulated and published'.

By the interwar period, additional federal legislation also gave the AEA a unique opportunity to define itself. For example, passed by the Sixty-seventh Congress in 1923, the Classification Act provided for the categorization and grading of technical and professional employees in the civilian branches of the federal government. Like their counterparts in many other fields, the leaders of the American Economic Association succeeded in linking this particular federal effort to their own continuing pursuit of professional cultivation. An early 1924 resolution of the AEA Executive Committee began steps to 'secure the classification of the technical economists in the professional and scientific services' of the federal government. The findings of a committee tasked to collate the results of this survey were reported to the Personnel Classification Board (of the US Civil Service Commission), the Committees on the Civil Service of the two houses of the Congress, and to the Executive Office of the President. In many respects the classification survey powerfully resonated with what had begun a decade earlier as part of the effort to support national mobilization for war. Yet here, in peacetime, it extended beyond the confines of an emergency canvass and became instead the basis of a continuing and ever more specific detailing of economics subspecialties. Indeed, for some older members of the profession the steps taken to stipulate as precisely as possible the expertise of individual practitioners could at times appear to narrow, and thereby adulterate, what the discipline as a whole had to offer. For most colleagues, however, that governmental needs melded so well with professionalizing strategies was cause for satisfaction rather than regret.

By the late 1930s, a segment of the AEA membership dissatisfied with the Association's perceived lack of attention to financial issues worked to create the American Finance Association (AFA). At the 1939 AEA Annual Meeting, the formal steps were taken to create the AFA. Although the Second World War slowed the evolution of the new organization, by 1942 the new journal *American Finance* appeared. It ultimately evolved into the well-known *Journal of Finance* just after war's end. Over 1,000 members populated the AFA ranks by the early 1950s.

In so far as a desire to distil professional opinion dated back to the early years of the Association's founding, it is not surprising to find that renewed interest along these lines emerged as economists turned their attention to planning for another war and its aftermath, and anticipating the role of economists in government during peacetime. During the Second World War the AEA leadership began deliberations 'to [consider ways of] making the informed opinion of our membership more effective in matters of public policy'. Because the Association, by the terms of its charter, could take no partisan positions, the trio nevertheless believed that the 'technical competence' of members could be expressed on 'matters of public importance'. This would require of course that 'all academically respectable views on

any posed controversial question be represented' on committees formed to pronounce on policy matters.

While striving to adhere to its strictures against partisan endorsements, a task made all the more difficult in the highly charged politics of the immediate post-war era, the leadership of the American Economic Association turned its attention to engagement with seemingly more 'objective' needs of the national security state. In these efforts, their work was paralleled by that of colleagues already assigned to some of Washington's highest echelons. Over the course of the 1950s, for example, government economists made frequent visits to the military service academies, and to such institutions as the War College of the Air Force and the Industrial College of the Armed Forces (of the National Defense University, Fort McNair, Washington, DC) to discuss (and participate in conferences on) such matters as 'mobilization of the national economy in the face of atomic attack', 'economic stabilization after attack' and 'domestic economies and their relation to national power'.

AEA officials also worked closely with colleagues on government duty to assist the national service academies in fully integrating an increasingly rigorous and operational discipline within their curricula. On behalf of the Armed Forces Institute, Secretary-Treasurer James Washington Bell coordinated the efforts of several scholars to oversee textbook selections in the field for cadets and midshipmen, thus 'prov[iding] the Armed Forces of the United States with educational materials which [we]re in accord with the best civilian practices' in economics as a whole. By the mid-1950s it had also become common for AEA functionaries to help designate particular professionals for work in special seminars on international organization and security convened by the transnational diplomatic and military alliance known as the North Atlantic Treaty Organization (NATO). It was a short step from these activities to involvement with the recruitment of undergraduate and graduate economics students for work within the now greatly expanded domain of the national security apparatus – including the Central Intelligence Agency (CIA).

The Post-War and Cold War Eras

Post-war reconstruction also brought the Association into the business of aiding professionals in devastated areas overseas. In addition to contributing free books and copies of the *American Economic Review* along with cash donations to scholarly libraries in Europe and East Asia, the AEA became involved in the revision of curricula and the rehabilitation and vetting of foreign faculties. American economists going overseas, on either official or personal tours, were asked by government authorities to check up on colleagues who had perhaps been imprisoned, wounded or otherwise victimized by German national socialism or Japanese imperialism. Letters to Association members from economists abroad often contained information regarding colleagues who either had or had not collaborated with the enemy. Efforts were made to raise money for the relief of those who had opposed fascism and militarism. A note from a German colleague to former AEA President Paul Douglas was forwarded to the Association offices because in it there was 'a very valuable list of economists who either opposed Hitler or kept their honor clean'. American economists were now in a position not only to secure greater influence and prestige at home but also to reconstitute virtually from scratch the European and Asian branches of the guild.

The reconstruction of foreign scholarly libraries prompted the American Library Association (ALA) to ask professional societies to provide book lists in their fields to guide rebuilding efforts. AEA officials canvassed the membership for suggestions and ultimately provided such lists, with regard to economics, to the ALA. With such recommended titles as *Stalin, A Critical Survey of Bolshevism* and *Marxism*: *An Autopsy,* the ideological content of the library aid effort seems clear. This is of course hardly surprising. The point here is not that American economists would generally be loath to suggest books that extolled Marxism or Stalin – indeed, AEA members and the AEA leadership utterly failed to defend beleaguered colleagues victimized by the anti-communist hysteria stoked by McCarthyism – but that Allied victory had the added impact of giving them a

great deal of influence on the future course of foreign scholarship in the field. If post-war reconstruction served to recast Europe and Asia in America's image, as some scholars have suggested, the representations of that process in the academic and intellectual world should not be overlooked.

Participation of the American economics profession in the emergent Pax Americana of the 1950s also expressed itself in a continuation and evolution of links between economists and the military–industrial establishment that had necessarily arisen in the 1940s. Economists of course participated both in the private sector and at the government level in the mobilization and allocation of resources for war. In addition, the profession became increasingly involved in establishing curricula at the nation's armed service academies on the economics of national security and defence. Defence-related research and support of basic economics investigations by armed forces agencies became more and more common. Moreover, the emergence of wholly new aspects of the discipline – such as 'linear programming' and 'input-output analysis' – was inherent in the association of professional economics with the national security state. The AEA even helped the U.S. Information Agency in securing prominent and competent personnel to do radio broadcasts on economic subjects for the Voice of America.

Curriculum revision and reform was a project that lasted well into the 1950s. Two months before the opening of a second front in western Europe the Association Executive Committee asked that the new Committee on Undergraduate Teaching and the Training of Economists concern itself with 'the long-run postwar period'. Ultimately, of particular interest to this committee with regard to the matter of undergraduate instruction were 'problems of indoctrination [of students] as to social consciousness and professional responsibility'. Four months after the surrender of Japan, 160 college and university economics departments around the country received questionnaires from the AEA soliciting information on undergraduate instruction. By the autumn of 1950 the AEA secretariat initiated plans for a conference on social science teaching at the pre-collegiate and collegiate levels. At the same time, the Committee on Graduate Training in Economics began its work, seeking to formalize in detail the professional requirements for the Ph.D. degree. To this effort, the Rockefeller Foundation donated $16,000. When the committee transmitted its findings to university deans and presidents, return correspondence was grateful and enthusiastic. War-related agendas thus carried over into long-standing peacetime activities.

Interestingly enough, and not surprisingly, concerns with the content and delivery of economics curricula emerged directly from Second World War experience. Wartime efforts on behalf of the National Roster of Scientific and Specialized Personnel (NRSSP) had made the leadership of the American Economic Association both particularly sensitive and responsive to requests for information about the discipline and its specialists. Moving from a focus on calculating the profession's numbers and activities, as the NRSSP had requested, to a self-conscious assessment of teaching methods, course content, and educational performance standards was altogether understandable and clear-cut. AEA initiatives in this regard were only further stimulated by the desire of the Veterans Administration and related agencies to facilitate the re-entry of armed forces personnel to civilian life after the Second World War and the Korean conflict.

Defining what an economist was, and what he or she did for a living, was one thing; stipulating how an economist was to be trained, not to mention evaluating his or her professional skills, was something else. In a series of studies, the first of which was launched in 1949, with follow-ups taking place throughout the 1950s, AEA task forces conducted wide-ranging surveys of undergraduate and graduate curricula throughout the country. Of particular importance to these committees were the 'opinions of leaders in graduate training' in the field at the nation's foremost research institutions. Recognizing that '[t]he Association ha[d] a definite professional responsibility in this [regard]', the Ad Hoc Committee on Graduate Training in Economics made its first report to the AEA Executive Committee late in

1950. Determined to guide universities in the establishment and maintenance of 'good graduate program[s] in economics at various levels', the committee particularly encouraged institutions to improve standards for the selection of incoming students, articulate precise objectives for advanced study in the field, and vet subject matter and course content with a view towards the rigorous training of new colleagues. Specifically, the committee believed that the 'important tools' in all graduate economics instruction were 'mathematics, accounting, statistics, history, logic, scientific method, and foreign language'.

Not least of the historical forces that shaped the continuing evolution of the American economics profession in the latter half of the 20th century was the unique prosperity the nation enjoyed throughout the 1950s and 1960s. If the application of a new learning to the management of a 'mixed economy' provided an exceptional opportunity for social scientific expertise to demonstrate its rigour and effectiveness, the context within which that display took place set the terms of both its practice and its success. Having proved its mettle in the extraordinary years of world wars, and having continued to do so in the early stages of what would be an even longer cold war, modern economic theory was now deployed in an altogether novel exercise: the pursuit and maintenance of full employment growth in peacetime. That, owing to history itself, the national economy was singularly well positioned for sustained expansion in the post-war period made that task all the more tractable.

Unlike any other industrialized nation in the world at the time, the United States met the 1950s with an economy not only physically intact but also organizationally and technologically robust. The demographic echoes of war set the stage for an acceleration in the rate of population growth, while the labour market effects of demobilization surprisingly sparked a rise in wages and incomes. Rapid and profitable conversion to domestic production was further stimulated by foreign demand – most vividly and poignantly emanating from those regions most devastated by the war itself – for the products of American industry and agriculture. As for international finance, the nation stood as creditor virtually to the entire world, and the dollar, both by default and by a multilateral agreement first reached by the Allied nations at Bretton Woods, had become a kind of numeraire to a newly emergent system of global commerce. With no small justification, the 1950s and 1960s came to be regarded as a golden age of American capitalism.

The Era of the 'New Economics' and Beyond

Macroeconomic management, demanding under any circumstances, was made substantially easier for a post-war generation that found itself the beneficiaries of historical circumstance. Farm from solving the cruel puzzle of idle capacity and widespread unemployment that had characterized the Great Depression, and unlike the challenge to rationalize allocation and maximize production in the emergency of war, the task that lay before American economists by the mid-1950s was both more straightforward and less difficult. More straightforward because, thanks to both the 'Keynesian revolution' in economic thought and the policy experience derived from mobilization and war, the relationship between individual market behaviour and aggregate outcomes was finally subject to systematic understanding. Less difficult because, given the sturdy rebound of the economy in the wake of the Second World War, there existed both the confidence (most especially exemplified by the moderate rates of return in the markets for Treasury bills and other government obligations) and the means (most vividly represented by rising income tax receipts) to realize fiscal spending targets with a minimum of redistributive implications.

So optimistic were politicians and the vast majority of economists concerning the effectiveness of stabilization policy techniques that it became fashionable by the early 1960s to speak of the 'end of the business cycle' and of the ability of policymakers to 'fine-tune' macroeconomic performance. In the *Economic Report of the President, 1965*, President Lyndon Johnson made it clear that he 'd[id] not believe recessions [we]re

inevitable' (Council of Economic Advisers 1965, p. 10). Similarly, in what was arguably the most influential economics textbook ever published, Paul Samuelson (1972, p. 250) wrote that his colleagues 'kn[ew] how to use monetary and fiscal policy to keep any recessions that br[oke] out from snowballing into lasting chronic slumps'. He went on to claim that the business cycle was thus a thing of the past. Expert knowledge buttressed by a healthy and resilient economy could now make the periodic deprivation and hardship once believed to be the inevitable consequence of the cycle truly a thing of the past.

Cultivating a politics of aggregate productivity and a discourse about sustained prosperity was not solely the result of professional self-assurance and self-promotion, nor was it simply the manifestation of a particular politician's (or a particular party's) strategy to procure votes. The focus on growth and accumulation so characteristic of the new economics of the post-war era represented as well a transformation in the nation's political culture that had been in the making for decades. For 19th-century convictions regarding the probity of thrift and self-improvement, mid-20th-century Americans had swapped a fascination with, and a virtual anxiety about, the individuation and comfort associated with consumption. Production was no longer an end in itself, nor could it alone provide meaning and dignity to one's life. Rather, it was the goods and services of the material world that afforded freedom and amenities, setting one's self off from others and liberating all from both the overt and the hidden injuries of class, ethnicity, and gender. What came to be known as the 'economic growthmanship' practised by a new social scientific elite was, on the one side, a particular aspect of a stage in the evolution of a professional community; on the other, it distilled, within a set of seemingly unassailable aspirations and beliefs, a society's unself-conscious embrace of an altogether new set of cultural ideals.

Within an economics of abundance and stability rested the ingredients of a prosperous commonwealth devoid of the class antagonisms and struggles over normative values that were a threat to both the legitimacy of social scientific policymaking and social tranquility and political cohesion. If an ' emphasis on an ever-growing pie, rather than on slicing up a given pie in a new way, [wa]s well designed . . . to attract widespread support' for particular policies (Tobin 1966, p. 42), it was also true that the depiction of the economy as a kind of positive-sum game from which all could benefit independent of their *relative* shares in particular outcomes was an essential part of the political-economic ideology of post-war America from the time of Truman's Fair Deal through that of Lyndon Johnson's Great Society, up to and including the early stages of Richard Nixon's New Federalism. Their specific analytical differences aside, virtually all mainstream American economists both embraced and relied upon this 'depoliticization' of the marketplace in their determination to separate positive economic 'science' from normative assertions. So long as the profession could retain this image of its work as a calculation of optimal means to a given end rather than the comparison of different and possibly incompatible goals, its claims to the authority and influence devoutly sought since the late 1890s were secure. As soon as that archetype was jettisoned or challenged, modern economics would find itself in a world, not of rigour and logic, but rather of ideological belief and political power.

Indeed, in December 1968 the Union for Radical Political Economics (URPE) held its first national conference in Philadelphia. This was done in opposition to the AEA's Annual Meeting in Chicago, which URPE interpreted as an endorsement of that city's violent response to anti-war demonstrations that summer. The AEA Executive Committee, chaired by then AEA President Kenneth Boulding, concluded that moving the Meeting would have violated the Association's policy of political neutrality. A year later, an activist disrupted the AEA Annual Meeting by reading a statement, at a plenary session, denouncing the Association for 'perpetuating professionalism, elitism, and petty irrelevance'. This led to a mass walk of 'radical economists'. In partial response to these insurgencies from within the ranks, the AEA established a Committee on the Status of Minority Groups in the Economics Profession (CSMGEP) in 1968 – and, by 1971, a Committee on the Status of Women in

the Economics Profession (CSWEP) and a working group on the status of minorities. The social change and turmoil of American society in the Vietnam War era had come home to the AEA itself.

In the mid-1980s, concerns regarding the training of new generations of economists came to the fore in AEA deliberations. At a National Science Foundation symposium held late in 1986, many participants argued that graduate curricula in economics had become exceedingly esoteric and abstract, of little use in the resolution of contemporary economic problems. A Commission on Graduate Education in Economics (COGEE) was subsequently charged to study the problem. It issued a report in 1991 that identified a number of problems in the profession such as a lack of focus on the inculcation of applied research skills, untoward emphasis on mathematics and axiomatic reasoning instead of analysing institutions and historical change, inadequate attention to the training with respect to communication and writing skills, an absence of creativity, and excessive emphasis on conformity and homogeneity in professional discourse. The COGEE report was so controversial that it was never accepted as an official AEA document.

Over a century ago American scholars eager to understand the economic world in which they lived embraced a project of both theoretical and social import. In doing so, they yoked the insights of an intellectual revolution in the ways social scientists understood human behaviour in commercial settings to a specific agenda of professional advancement. A late-19th-century transformation in economic thought afforded these investigators a powerful and versatile set of tools with which to situate human rationality at the centre of a remarkable and immensely influential human institution – the marketplace. A 'science' of individual behaviour and social organization was thus established, the implications of which played no small part in the creation of a respected and ultimately quite accomplished community of professional experts – as exemplified by the AEA.

But an authoritative community does not, precisely because it cannot, subsist on its own. American economists were most eager to place their skills at the service of the state. Here history proved both a blessing and a curse, for the profession's great achievements of the 20th century, especially but not solely during years of global conflict and war, were also paralleled by failures and betrayals emanating from the same source. Indeed, it would be these negative moments in the century-long progress of their self-realization that would drive economists and their discipline farther and farther from engagement with the affairs of state in favour of an increasingly introverted and surprisingly opaque discourse. At the same time, eager like most professionals to retain an influence and visibility in public affairs that would cultivate a continued appreciation of their virtues and skills, later generations of economists would make themselves – whether consciously or not – useful servants of those, in both the political and the commercial worlds, who had an altogether different view of public purpose and of the appropriate role of government.

Bibliography

AEA (American Economic Association). n.d. *Records of the American Economic Association*. Evanston: Archives, Northwestern University Library and Nashville: Secretary's Office, Vanderbilt University.

Bernstein, M.A. 2001. *A perilous progress: Economists and public purpose in twentieth-century America*. Princeton: Princeton University Press.

Coats, A.W. 1960. The first two decades of the American Economic Association. *American Economic Review* 50: 555–574.

Coats, A.W. 1964. The American Economic Association, 1904–1929. *American Economic Review* 54: 261–285.

Coats, A.W. 1985. The American Economic Association and the economics profession. *Journal of Economic Literature* 23: 1697–1728.

Coats, A.W. 1993. *The sociology and professionalization of economics,* 2 vols. New York: Routledge.

Council of Economic Advisers. 1965. *Economic report of the president, 1965*. Washington, DC: U.S. Government Printing Office.

Dorfman, J. 1936. The founding and early history of the American Economic Association. *American Economic Review* 26: 141–150.

Dorfman, J. 1949, 1959. *The economic mind in American civilization*, vols. 3–5. New York: Viking Press.

Furner, M.O. 1975. *Advocacy and objectivity: A crisis in the professionalization of American social science, 1865–1905*. Lexington: University Press of Kentucky.

Ross, D. 1991. *The origins of American social science*. New York: Cambridge University Press.

Samuelson, P.A. 1972. *Economics: An introductory analysis*. New York: McGraw-Hill.
Tobin, J. 1966. *National economic policy: Essays*. New Haven: Yale University Press.
U.S. Council of Economic Advisers. 1965. *Economic report of the president: 1965*. Washington, DC: US Government Printing Office.

American Exceptionalism

Louise C. Keely

Abstract

'American exceptionalism' refers to significant differences between the United States and Western Europe, first identified by European commentators in the 19th century, including the circumstances surrounding the founding and settlement of the United States, as well as a concept of nationhood based on immigration rather than a shared history. Economists have contributed in important ways to the documentation and evaluation of exceptionalism's economic effects. One important example of this research is on contemporary differences in social policy between the United States and Western Europe.

Keywords

American exceptionalism; Crime; De tocqueville, a.; Equality of Opportunity; Fertility; Higher Education; Individualism; Inequality; International Migration; Productivity; Racism; Redistribution; Religion; Slavery; Socialism; Trade Unions; Voluntarism; Western Europe; Workfare

JEL Classifications

D6; D7; E6; H00; J00; P52; E61; H11

The term 'American exceptionalism', which has been current among scholars since Alexis de Tocqueville coined it, captures the idea that America is different in important ways from Western European countries. This exceptionalism is, at first glance, surprising given that the United States was initially settled and governed by persons from Europe and that in many ways the two regions appear relatively similar. The term suggests a set of reasons for the differences in institutions and individual choices in the realms of politics, economics, and social interactions. (For some insightful and broad discussions of American exceptionalism, see Lipset 1996; Shafer 1999.)

Comparing America with Western Europe is somewhat arbitrary. The attention paid to American exceptionalism does not suggest that other countries are not also exceptional. Indeed, other examples of exceptionalism have been studied by social scientists.

Interest in the United States is due in part to its economic and military power. Simply put, the US government and economy exert a significant influence on all countries, including those of Western Europe. But there is an historical reason for the comparison with Western Europe. Europeans settled and governed the region that became the United States of America. It was Europeans who in the 19th century visited and wrote about the United States, comparing it with their native lands. De Toqueville is the best known, but he is only one of several Europeans who were interested in what they saw as a profound contrast between the United States and Western Europe.

It is useful to define American exceptionalism in terms of origins rather than consequences. Political, economic and cultural outcomes, whether observed today or in the 19th century, are endogenous. However, there may be circumstances distinguishing the United States from Western Europe that can be treated as fundamental, or exogenous, to the United States as a sovereign state. Those circumstances may have led to the differences in outcomes observed in the 19th century and today.

Before the American Revolution that began in 1776, the British governed the colonies that came to constitute the original United States. The constitution of the United States can be understood as a product of both the trauma of the revolution and the fact that 13 geographical areas, with distinct identities, were creating a single federal government. Moreover, the framers of the constitution were themselves diverse not only in place of

origin but also in social and economic background (Mee 1987). The constitution contains features reflecting a certain distrust of centralized public authority. The increase in popular political participation beyond that which existed under colonial administration, the checks and balances across the three branches of government, and the restrictions on the powers of the federal government are prominent indications of this concern.

Europeans, many with a specific religious agenda, initially settled the area that became the United States. They aimed to create a society directed by divine providence. These settlers faced unusual circumstances in modern history, having the opportunity not only to establish a government largely from scratch but also to settle a large geographic area that was either uninhabited or inhabited by people they could displace, albeit sometimes with difficulty.

The historical circumstance of the United States as a state whose citizens' families came from other countries within recent memory led, in part, to a notion of nationality that was flexible from the beginning. What it is to be American has never, with the important exception of slaves who were not treated as full citizens, been dependent on ethnic background or common historical circumstances. This is not to deny that racism or ethnic prejudice have existed in the United States, but rather to say that what it is to be *American* has never been predicated on a particular origin or history.

This notion of nationality lent itself to the United States' openness to immigrants from many countries, until recently mostly Europeans. Some immigrants came to the United States to escape political or religious persecution, such as the Jews during the pogroms of the late 19th century and during and after the fascist regimes that held sway in Europe in the first half of the 20th century. However, many more came in pursuit of economic opportunity. Some immigrants, such as the Irish in the mid-19th century, faced terrible economic circumstances in their home countries. Others chose to emigrate under less dire constraints.

Across these diverse circumstances of immigration, it is generally the case that immigrants to the United States were self-selected into this group. The important exception to this self-selection is the immigrants from Africa and the Caribbean who were brought as slaves to the United States.

While self-selecting immigrants left their countries of origin for a variety of reasons, they would all have believed that in the United States their lives would be better in economic, political or religious terms. By seizing the opportunity to become American, they could lead better lives (loosely defined). This possibility is attributable to exogenous circumstances: the physical expansiveness of the United States and the related expandable notion of American nationality. The populations who chose to move to the United States also did so in large part because they believed that self-determination was possible in the United States.

Such self-determination is part of the ideology on which (rather than on a common history) the United States was founded, and subscription to which makes one American. That ideology includes a set of values and institutions that are immediately familiar as distinctly American. Americans are viewed (and are thought to view themselves) as relatively distrustful of public authority and as embracing self-reliance. Broadly speaking, they subscribe to the ideals of equal socio-economic opportunity (as distinct from equality in outcomes), a classless society, and an inclusive democratic process. American institutions are relatively fragmented and public services are generally viewed to be less comprehensive than in countries with similar per capita incomes. Americans are more religious than Europeans. The concept of American nationality is relatively inclusive.

Why should American exceptionalism matter to economists? What role does economics have to play in understanding the consequences of American exceptionalism?

At least three distinct avenues of enquiry are of interest to economists. The first is positive: to document outcomes that may be attributable to American exceptionalism. The second is evaluative: to examine whether the exceptional circumstances under which the United States and its citizenry were constituted have led to differences between both the institutions and the values and beliefs (or culture) of Americans and those of Western

Europeans. (A substantial political science literature debates the relative importance of institutional differences and cultural differences in defining American exceptionalism. The present author finds that discussion unclear, and thinks that it is more useful to view both types of differences are outcomes of exceptionalism rather than manifestations of it. That is, both types of differences may exist and are not mutually exclusive.) The third avenue of enquiry is normative: given evidence of exceptionalism, the task is to examine the context in which economic policies in the United States are to be designed and evaluated relative to Western Europe.

Existing research focuses on American-European differences in political, cultural, and economic outcomes, and asks questions including those in the following non-exhaustive list:

- Why was there not a socialist movement in the United States? (Jacoby 1991; Lipset and Marks 2000; Voss 1993)
- Why have labour unions been weaker in the United States than in Western Europe? (Currie and Ferrie 1995; Freeman 1994; Jacoby 1991; Voss 1993)
- Why do Americans publicly redistribute income less than Europeans do? (Alesina and Glaeser 2004; Benabou and Tirole 2004; Shafer 1991)
- Why do Americans perceive a higher probability of socio-economic mobility within and across generations than those in Western Europe? (Keely 2005a)
- Why is the US higher education system larger than those in European countries? (Shafer 1991)
- Why is there more violent crime in the United States? (Shafer 1991)
- Why is productivity in the United States higher than in Western Europe? (Abramoviz and David, 1994; Gordon 2002; Romano 1993)
- Why do Americans participate in volunteer activity more than Western Europeans? (Lipset in Shafer 1991; Lipset 1996)
- Why did the institution of slavery persist in the United States long after it disappeared from Western Europe? (Shafer 1991)
- Why are Americans more religiously observant than Western Europeans? (Shafer 1991)
- Why is fertility higher in the United States than in Western Europe? (Keely 2004, 2005b)
- Why has the United States been able to assimilate immigrants at levels well beyond those of Western Europe? (Glazer 1999)

Proposed answers to these questions have one common element: American exceptionalism. These issues are all directly or indirectly related to economic policy, and pose questions that economists' tools can help to answer. Consider for example the question: Why do Americans publicly redistribute income less than Europeans do? Economists have recently tried to answer this question.

A first step is to document the differences in redistribution. OECD data indicate that, while public spending on social services amounted on average to 24 per cent of GDP in Western European countries, in the United States it amounted to 15 per cent. In the United States private social spending as a share of the total in 1995 is reported by the OECD to be 41 per cent, while for European Union countries it varied from per cent (Spain) to 16.9 per cent (the United Kingdom) (OECD 2005).

Second, how can this difference be attributed to American exceptionalism rather than to some other source? Identifying the effects of exceptionalism as such is extremely difficult. Competing hypotheses about the same outcome can be observationally equivalent. However, the models that lead to the same predicted outcome may also contain secondary predictions that do vary across models. That variation may be exploited to compare hypotheses. One suggested approach has been predicated on the higher level of ethnic heterogeneity in the United States than in Western Europe. The institution of slavery, which led to the existence of a minority of citizens of African origin, and the flow of ethnically varied immigrants into the United States have been attributed to American exceptionalism.

Heterogeneity itself doesn't explain why there is less income redistribution. Some authors have proposed that heterogeneity may matter in terms of its interaction with preferences.

(This hypothesis has been proposed by Alesina et al. 1999, and Luttmer 2001. See Keely and Tan 2005, for related discussion.) The assumption regarding preferences is that agents experience disutility when they observe people who differ from them in some salient dimension such as race to be more likely recipients of public income redistribution. Such preferences capture a notion of racism.

Racism is not a feature or direct consequence of American exceptionalism as I have defined it. Nor are norms regarding interracial interactions exogenous or unchanging variables. Interactions between, and socio-economic outcomes across, racial and ethnic groups in the United States have changed enormously (though perhaps still not enough) over the past century. Therefore, this preference-based hypothesis regarding different levels of income redistribution in the United States and Europe is only partially based on an observation directly attributable to American exceptionalism.

An alternative hypothesis that relies more squarely on exceptionalism, rather than on other cultural or political assumptions, is as follows. People face uncertainty about future income and whether they will be net beneficiaries of income redistribution policies. In order to form the expectations that are necessary to determine preferred income redistribution policy, people may use information about others who are similar to them in ways that are relevant to income determination. In a society where racial or ethnic characteristics are correlated with income, race and ethnicity can be a factor determining similarity. If the size of the minority group (in this case, blacks) is sufficiently large and/or the difference in the groups' income distributions is sufficiently large (in some well-defined way), then it can be the case that whites, who have higher average income, are less likely to be in favour of income redistribution than are blacks.

This hypothesis relies on three factors that have been traced directly to American exceptionalism: (*a*) the ethnic heterogeneity of agents; (*b*) income inequality linked to the legacy of the institution of slavery (given the presence of relatively large amounts of arable land); and (*c*) the focus on individualism rather than communal obligation.

Both hypotheses lead to a prediction that the United States has lower levels of redistribution than Western European countries. How can competing hypotheses be evaluated? As suggested above, one strategy is to look for secondary and testable predictions that differ across hypotheses. While there is a history in the United States of racism connected to whites and blacks, there is also a history of racism against other ethnic groups such as Asians and Hispanics. Certainly there is a widely recognized ethnic distinction between those groups on the one hand and people of European descent on the other. If differences in income redistribution preferences are due to racism, then it should be the case that exposure to ethnic heterogeneity of these types should also lead to stronger opposition to redistribution overall.

In contrast, if differences in redistribution preferences stem from differences in income distributions conditional on ethnic group, then an effect of heterogeneity might not be uniform. For instance, if the conditional distribution of whites and Asians is not statistically significantly different, then income redistribution preferences are predicted to be lower in areas with more heterogeneity in the white-Asian dimension only under the first 'racism' hypothesis.

The third way in which American exceptionalism matters to economists is its impact on political economy parameters. Every public authority is policy constrained, for instance by cultural values and economic circumstances. America was founded on an ideology that, it has been argued, persists. While its details and interpretation may change, its essence is constant. Any normative statement regarding the political economy of the United States should, in the face of strong evidence of American exceptionalism, take account of those constraints. More specifically, one of the ways in which American exceptionalism manifests itself and has been summarized is the claim that individualism and anti-statism lead to a notion of egalitarianism based on opportunity rather than outcomes.

In this light, it is completely unsurprising that the United States has a smaller welfare state than those of Western Europe. Moreover, the types of welfare reform that have been instituted since 1995 and the rhetoric used to promote them are also consistent with American exceptionalism. Welfare is now sometimes called workfare; there is a push to move welfare towards a policy that provides opportunity through job training and work rather than providing a guaranteed outcome through direct transfers. Private involvement in a publicly administered welfare programme also seems more politically feasible than a purely public model as in Western Europe.

American exceptionalism is an old idea. In his now famous 1630 'City on a Hill' speech, John Winthrop spoke thus of the newly settled land:

> [W]ee shall finde that the God of Israell is among us, when tenn of us shall be able to resist a thousand of our enemies, when hee shall make us a prayse and glory, that men shall say of succeeding plantacions: the lord make it like that of New England: for wee must Consider that wee shall be as a Citty upon a Hill, the eies of all people are uppon us. . . .

Economists have a perspective and set of skills to contribute towards understanding the extent to which American exceptionalism exists and its implications for Americans and people in other countries.

See Also

► Equality of Opportunity

Bibliography

Abramovitz, M., and P. David. 1995. Convergence and deferred catch-up: Productivity leadership and the waning of American exceptionalism. In *Growth and development: The economics of the 21st century*, ed. R. Landau, T. Taylor, and G. Wright. Stanford: Stanford University Press.

Alesina, A., and E. Glaeser. 2004. *Fighting poverty in the US and Europe: A World of difference*. Oxford: Oxford University Press.

Alesina, Al, R. Baqir, and W. Easterly. 1999. Public goods and ethnic divisions. *Quarterly Journal of Economics* 114: 1243–1284.

Benabou, R., and J. Tirole. 2004. *Belief in a just world and redistributive politics. Working paper*. Princeton: Princeton University.

Currie, J., and J. Ferrie. 1995. *Strikes and the law in the US, 1881–1884: New evidence on the origins of American exceptionalism*. UCLA: Working paper.

Freeman, R. 1994. American exceptionalism in the labor market: Union-nonunion differentials in the United States and other countries. In *Labor economics and industrial relations: Markets and institutions*, ed. C. Kerr and P. Staudohar. Cambridge, MA: Harvard University Press.

Glazer, N. 1999. Multiculturalism and American exceptionalism. In *Multicultural questions*, ed. C. Joppke and S. Lukes. Oxford: Oxford University Press.

Gordon, R. 2002. *Technology and economic performance in the American economy*, Working Paper No. 8771. Cambridge, MA: NBER.

Jacoby, S. 1991. American exceptionalism revisited: The importance of management. In *Masters to managers: Historical and comparative perspectives on American employers*, ed. S. Jacoby. New York: Columbia University Press.

Keely, L. 2004. Perceived return to human capital in fertility. Working paper. University of Wisconsin-Madison.

Keely, L. 2005a. Mobility beliefs and socioeconomic mobility. Working paper. University of Wisconsin-Madison.

Keely, L. 2005b. Intergenerational socioeconomic mobility and fertility: A new look at an old question. Working paper. University of Wisconsin-Madison.

Keely, L., and C.M. Tan. 2005. *Understanding preferences for income redistribution. Working paper*. Medford: Tufts University.

Lipset, S. 1996. *American exceptionalism: A double-edged sword*. New York: W.W. Norton.

Lipset, S., and G. Marks. 2000. *It didn't happen here: Why socialism failed in the United States*. New York: W.W. Norton.

Luttmer, E. 2001. Group loyalty and the taste for redistribution. *Journal of Political Economy* 109: 500–528.

Mee, C. 1987. *The genius of the people*. New York: Harper and Row.

OECD (Organisation for Economic Co-operation and Development). 2005. *Social expenditure database*. Paris: OECD.

Romano, R. 1993. *The genius of American corporate Law*. Washington, DC: AEI Press.

Shafer, B., eds. 1991. *Is America different?* Oxford: Oxford University Press.

Shafer, B. 1999. American exceptionalism. *Annual Review of Political Science* 2: 445–463.

Voss, K. 1993. *The making of American exceptionalism: The knights of labor and class formation in the nineteenth century*. Ithaca: Cornell University Press.

Winthrop, J. 1639. City upon a Hill. Online. Available at http://www.mtholyoke.edu/acad/intrel/winthrop.htm. Accessed 30 Aug 2005.

Amoroso, Luigi (1886–1965)

Giancarlo Gandolfo

Keywords
Amoroso, Luigi; Consumer equilibrium; Pareto efficiency

A mathematician by training (at the Normale, Pisa), Amoroso was assistant professor of mathematics in Rome, then professor of financial mathematics in Bari, but soon turned to economics, which he taught from 1921 in Naples and then Rome. He was a fellow of the Econometric Society.

Leaving aside his contributions to pure mathematics (e.g. 1910), financial mathematics (e.g. 1921a), statistics (e.g. 1916), demography (e.g. 1929), four books (1921b, 1938, 1942, 1949) well summarize his contributions to economics, also contained in over 100 articles.

Inspired by Pareto, his mathematical background led him to develop the analogy between pure economics and classical mechanics: the principle of minimum (use of scarce) means is the equivalent of the principle of least action. He also saw analogies between Heisenberg's uncertainty principle and economic phenomena, but did not develop this idea. His existence and uniqueness proof (1928) of a meaningful solution to the system of equations defining consumers' equilibrium is the first modern treatment of existence and uniqueness problems in economics.

Amoroso stressed the need to analyse all optimum conditions in a dynamic context: for example, the consumer maximizes a function under the balance constraint expressed as a differential equation; the problem is solved by applying the calculus of variations. He thus derived the extension of Pareto's static optimum conditions to a dynamic context. By considering the market determination of prices and introducing relationships between inventories and prices, he obtained systems of integro-differential equations capable of causing cycles around a trend, thus giving an explanation for crises and secular movements.

Selected Works

A full bibliography of Amoroso's works up to 1959 and an evaluation of his scientific contributions by various authors is contained in: Onoranze al Prof. Luigi Amoroso, Annali dell'Istituto di Statistica, vol. 30, Università di Bari, 1959.

1910. Sulla risolubilità della equazione lineare integrale di prima specie. *Rendiconti della Accademia dei Lincei*. Classe di Scienze fisiche, matematiche e naturali. Nota presentata dal socio Castelnuovo, nella seduta del 16.1.1910, Roma.

1916. Contributo al metodo delle minime differenze. *Giornale degli Economisti e Rivista di Statistica*: 50–86.

1921a. *Lezioni di matematica finanziaria*, vol. 1 (1921), vol. 2 (1923). Naples: Majo.

1921b. *Lezioni di economia matematica*. Bologna: Zanichelli.

1928. Discussione del sistema di equazioni che definiscono l'equilibrio del consumatore. *Annali di Economia*: 31–41.

1929. L'equazione differenzialle del movimento della popolazione. *Rivista Italiana di Statistica*: 151–157.

1935. La dynamique de la circulation. *Econometrica* 3(4): 400–410.

1938. *Principi di economica corporativa*. Bologna: Zanichelli.

1942. *Meccanica economica*. Città di Castello: Macri.

1949. *Economia di mercato*. Bologna: Zuffi.

Amortization

Charles R. Hulten

Keywords
Amortization; Balloon loan; Depreciation; Investment; Asset value

JEL Classifications
M4; G00; M40

'Amortization' is an accounting term meaning the allocation of a cost to several time periods. The term is derived from the Latin word for 'death' and literally means to 'kill off' the liability. Debts which are paid off gradually are said to be amortized.

The term is also applied to the depreciation costs of the cost of certain assets which are used up in producing income. Amortization in this second sense is illustrated by the following example (Table 1). A firm spends $10,000 to invent and patent a new product which is expected to yield revenue (net of operating expenses) of $5,000 in the first year of production, $2,000 in each of the next three years, and $1,500 in the fifth year (see column (3) of Table 1). The product is assumed to become obsolete at the end of five years and to generate no additional revenue. The patent thus becomes valueless at that time.

The present value of the net revenue stream associated with the invention is initially $10,000 at an approximate ten per cent rate of discount. However, the present value of the remaining net revenue falls to $6,000 at the end of the first year, to $4,599 at the end of the second year, to $3,058 and $,1364 at the end of the third and fourth years, and to zero at the end of the product's useful life (see column (4)). This implies that the original $10,000 investment has been eroded by $4,000 at the end of the first year, $1,401 in the second year, and so on (see column (5)). In considering how much profit is earned in the first year, the loss in the value of the investment must be subtracted from revenue in order to keep the original value of the investment intact. Thus, profit in the first year is $1,000, or ten per cent of the original investment. Inspection of columns (4) and (6) reveals that the ratio of profit to remaining present value in the previous year is always ten per cent.

If, on the other hand, the reduction in value is not recognized as a cost, one would erroneously conclude that the investment yielded $12,500 over the life of the asset (the sum of column (3)) rather than $2,500 (the sum of column (6)). However, the value of the investment would have fallen from $10,000 to zero. To avoid a misstatement of profit for tax and financial accounting purposes, investors are allowed to amortize the cost of the asset over its useful life. A pattern of amortization that matches the actual yearly loss in asset value is usually termed 'economic depreciation', although this typically (but not always) applies to tangible capital like plant and equipment, while 'amortization' is often used in the context of intangible assets. The actual loss in value is often hard to measure and, in practice, reasonable assumptions about useful asset life and about the pattern of value loss are used (for example, the straight-line and declining-balance patterns).

The graduation write-off of a debt is another context in which the term 'amortization' is frequently used. The level-payment home mortgage is, for example, a common type of amortized loan. In the level-payment mortgage, the sum of the interest and principal payments is constant. During the early life of the loan, the bulk of this constant (or 'level') payment is for interest on the outstanding balance of the loan. The proportion of the level payment allocated to the

Amortization, Table 1 Amortization of hypothetical asset

(1) End of:	(2) Outlay	(3) Net revenue	(4) Present value*	(5) Loss in value	(6) Profit
yr 0	$10,000	0	$10,000	0	0
yr 1	0	$5,000	$6,000	$4,000	$1,000
yr 2	0	$2,000	$4,599	$1,401	$599
yr 3	0	$2,000	$3,058	$1,541	$459
yr 4	0	$2,000	$1,364	$1,694	$306
yr 5	0	$1,500	0	$1,364	$136

*Present value of remaining net revenue calculated using discount rate of 9.992%

repayment of principal gradually increases as time goes by, since interest is paid on the outstanding balance of the loan. In the fully amortized loan, the sum of the period-by-period repayments of principal over the life of the loan is equal to the original value of the debt.

This type of arrangement may be contrasted with the case of the 'balloon' loan, in which the entire principal is repaid at the termination date of the loan. Loans may be a mixture of the two types: amortization of part of the principal with a balloon payment equal to the unamortized balance.

See Also

- ▶ Capital Measurement
- ▶ Depreciation

Analogy and Metaphor

Rom Harré

Keywords
Analogy; Metaphor; Models; Simile

JEL Classifications
B4

We say that something A is analogous to something B if, in some relevant respect, A is similar to but not identical with B. This is the basic relation upon which the use of analogy in various kinds of reasoning depends. We speak of reasoning by analogy when on the basis of some similarity which we discern between two things or processes or properties, or what you will, we infer some other similarity. Reasoning by analogy is a special case of inductive reasoning since we must be wary of the possibility that the further similarities which are presupposed in our inference may not actually obtain. Like all inductive inference reasoning by analogy is stepping from the known to the unknown. Clearly, then, analogical reasoning is not demonstrative or deductive.

A more refined analysis of the structure of the analogy can be made by distinguishing between those respects in which the analogues are similar, called the positive analogy, those respects in which they are different, called the negative analogy, and those respects in which we are unsure whether the property in question marks a similarity or a difference – the neutral analogy (Hesse 1963). Once we have introduced the idea of neutral analogy the relation between the analogues is no longer symmetrical. If we think of analogy simply in terms of similarities and differences then if A is similar to B, B is similar to A, and if A is dissimilar to B, B is dissimilar to A. It does not matter which of A and B we say is analogous to which. But once we introduce the idea of neutral analogy we are obliged to decide which of the items under comparison is the one from which our reasoning will take a start and usually this decision is dependent on which of A or B we are confident we know. For example, if we argue that an illness is analogous to the invasion of a country by a hostile army, as van Helmont proposed in the 17th century, it seems reasonable to take the invasion by the hostile army as the term about which we can in principle know a great deal and the cause of illness as the term about whose properties we are less certain. In reasoning by analogy, then, about the cause of disease, the idea of an invasion is the given term and the illness is the unknown. We can then take the known properties of invasions and armies and set out on an experimental programme to decide how many properties similar to them are to be found in the causes of disease. Thus: 'Soldiers are organisms', 'Are the causes of disease micro-organisms?' The logic of analogy then consists in picking out sets of properties and making comparisons between the members of the one set and of the other.

In judging the force of an analogy we must have some way of deciding which properties are important and which are not. If two things are similar only in unimportant or inessential ways and differ in other respects, then we generally take the analogy between them to be weak. Unlike deductive reasoning, analogy is, therefore, highly sensitive to context and to the interests of whoever

is making use of it. It can hardly be said that there is anything intrinsic about a property which makes it important. Rather its importance depends upon the context and interests of the user. Furthermore, we need also to assume that we can make some sort of quantitative assessment of the degrees of similarities and differences between the analogues and this may be quite difficult to do in any principled way.

I have described the relation of analogy in terms of concrete relations of similarity and difference between the properties of analogous things. However, there are important linguistic phenomena which are in some ways like an analogy. The most obvious is simile. When we use that figure of speech we explicitly invite a comparison between the referents of the terms between which the simile is drawn by reference to likenesses. We tacitly assume that we draw a simile only where there are also differences. There are plenty of literary examples to illustrate this relationship.

The analogy relation seems to have another realization in language in metaphor. In a metaphorical use of a term an expression is employed in a novel context. Words which are customarily used for discussing one kind of subject matter, are used to describe some other. Some have said that in metaphor the sense of a word is displaced. In order for a metaphor to have any bite it must reflect some similarity. The metaphor 'life's journey' would hardly have had the currency that it enjoys in improving discourses, such as the speeches which accompany school prize-givings, had there been no way in which life could be seen as a journey. But unlike simile, metaphorical uses do not leave words unaffected. It has been pointed out by many students of metaphor that when a concept is displaced into a new domain it not only serves to highlight some hitherto unnoticed similarity between its old and new referents, but it changes its significance through coming to be used in a new domain. So the term 'current' was first used in the description of electricity, to highlight similarities between electricity and more easily observable fluids. The two centuries of use of this term in the electrical domain have certainly led to a change in its meaning (Martin and Harré 1982).

Analogies and Models

The recent trend in philosophy of science to look more closely at actual examples of scientific reasoning has disclosed the quite central role that analogical reasoning plays in both the physical sciences and the social sciences. A special terminology has grown up in the sciences by which the term 'model' is appropriated for concrete analogues (Bunge 1973).

Scientific models are of two main kinds. There are heuristic or homoeomorphic models and explanatory or paramorphic models. Each kind has a specific use.

Many phenomena are too complicated for ready examination. Salient features can be brought out by abstracting a simpler form from the original complexity and idealizing its properties. A homoeomorphic or heuristic model is a convenient representation of its subject. It may be a concrete thing, such as the scale models used in engineering. But it may be an abstract conceptual representation embodied in something like the 'rational actor' assumption in economics. Heuristic models are conservative. In a sense they merely represent what we already know but in some useful or convenient form.

Explanatory models (paramorphic analogues) are used creatively. They enable scientists to conceive of new kinds of beings and so far unobserved processes. Their main use is to complete theories by standing in for unobserved and so currently unknown causal processes. The kinetic theory depends on the idea of a swarm of molecules which are a model or analogue of the unknown constitution of real gases. The hypothetical behaviour of the molecular analogue must be like (analogous to) the behaviour of the real gas. Such models are of great interest to methodologists since they not only form the core of most scientific theories, but are also the vehicles for much creative scientific thinking. They are not devised at random. Their construction is always controlled by some implicit metaphysical assumptions (in the gas model case Newtonian atomism) which ensure their plausibility to the scientific community. This means that they are balanced between two analogy relations. They must behave

analogously to the real thing they are a model for; and they are constructed by analogy with the real thing they are modelled on. For instance, the popular rule-following models in social psychology should replicate the behaviour of the unknown cognitive systems they are models for while they must lie within the constraints imposed by the real cases of rule-following, say in ceremonial action, which they are modelled on. Both analogy relations are usually open, that is, though they exhibit positive and negative aspects, similarities and differences, there is usually a degree of unexplored neutral analogy. Theories develop by the conceptual exploration and, in favourable cases, the empirical testing of the neutral analogy.

Explanatory and heuristic models can be neatly distinguished by reference to their constitutive analogies. For a heuristic model source and subject are identical. A model plane is a model of a plane. But for an explanatory model source and subject are distinct. The idea of an implicit rule is modelled on that of an explicit rule, but the former is an analogue of some unknown regulative cognitive process.

Bibliography

Bunge, M. 1973. *Method, model and matter*. Dordrecht: Reidel.

Hesse, M. 1963. *Models and analogies in science*. London: Sheed and Ward.

Martin, J., and R. Harré. 1982. Metaphor in science. In *Metaphor: Problems and perspectives*, ed. D. Miall. Brighton: Harvester Press.

Anarchism

George Woodcock

A doctrine whose nature is suggested by its name, derived from the Greek *an archos,* meaning 'no government'. The term *anarchist* appears to have been first used in a pejorative sense during the English Civil War, against the Levellers, one of whose enemies called them 'Switzerizing anarchists', and during the French Revolution by most parties in deriding those who stood to the left of them in the political spectrum. It was first used positively by the French writer Pierre-Joseph Proudhon in 1840 when, in his *Qu'est-ce-que la propriété?* (*What is Property?*), a controversial essay on the economic bases of society, he defined his own political position by declaring, perhaps to shock his readers into attention, 'I am an anarchist.' Proudhon then explained his view that the real laws by which society operates have nothing to do with authority but are inherent in the very nature of society; he looked forward to the dissolution of authority and the liberation of the natural social order which it submerged. He went on, in his rather paradoxical manner, to declare: 'As man seeks justice in equality, so society seeks order in anarchy. Anarchy – the absence of a sovereign – such is the form of government to which we are every day approximating.'

Proudhon's attitude was typical of the anarchists in all periods. They have argued that man is a naturally social being, who through mutual aid evolves voluntary social institutions that can work effectively without the need for government, which in fact inhibits and distorts them. The important transformation of society, anarchists argue, will not be the political one of a change of rulers or a change of constitution, since political organization must be discarded; it must be replaced by the economic organization of the resources of a society without government. Thus, while they differ from socialists and communists in denying the state and any form of state control or initiative, anarchists agree with them in being opposed to capitalism, in seeking to abolish what one of their earliest thinkers, William Godwin, called 'accumulated property' and to replace it with some kind of common ownership of the means of production. Only a few extreme individualists have stood outside this pattern, as Max Stirner did.

The basic ideas of anarchism predate the use of the title *anarchist*. Some historians have found their origin in early religious movements that stood outside ordinary society, refused to obey its laws and attempted in some way to own their

goods in common, like the Essenes, the Anabaptists and the Doukhobors. But in these cases the search seems to have been for spiritual salvation through a progressive retreat from involvement in the material world, and they have little in common with anarchism as a secular doctrine directed towards social transformation.

However, there are at least two social thinkers anterior to Proudhon who seem to fit the necessary criteria to be regarded as anarchists, since (a) they present a fundamental criticism of the existing governmental structure of society; (b) they present an alternative libertarian vision of a society based on cooperation rather than on coercion; and (c) they propose a method or methods of proceeding from one to the other.

The first is Gerrard Winstanley, the leader of the Diggers, a small communitarian group who emerged in England during the Commonwealth. In his 1649 pamphlet, *Truth Lifting Up its Head Above Scandals,* which departed entirely from religious orthodoxy by equating God with Reason, Winstanley laid down what afterwards became basic propositions among the anarchists: that power corrupts, that property and freedom are incompatible, and that authority and property between them are the main causes of crime; that only in a rulerless society where work and products are shared will men be both free and happy, because they will be acting according to their own judgements and not according to laws imposed from above. Winstanley went beyond theory to direct action when he declared that only by their own action could the people change their lot, and he led his own followers in an occupation of English common lands, where they sought to set up an agrarian community in which all goods were shared. Despite the passive resistance they offered, the Diggers were finally forced off their land and Winstanley vanished into obscurity.

His ideas lingered in the dissenting sects of the 18th century, where they were picked up by William Godwin. In 1793 he published a massive treatise on the nature of government, *Political Justice*, which has often been described as the most thorough exposition of anarchist theory, though Godwin never called himself an anarchist. *Political Justice* does in fact admirably present the classic anarchist arguments that authority is against nature and that social evil exists because men are not free to act according to reason; 'accumulated property' is to be condemned because it is a source of power over other men.

Godwin anticipated the general anarchist emphasis on decentralization by sketching out a social organization in which the small autonomous community, or parish, would be the basic unit. He envisaged a loose economic system in which he anticipated Marx's slogan, 'From each according to his abilities, to each according to his needs', by proposing that – capital in the form of 'accumulated property' having been dissolved – men would freely transfer goods to each other according to need, and all would share in production. Though he seems to have imagined fairly accurately the labour-saving powers of machinery, since he prophesied a drastic reduction of the work day, he does not appear to have taken into account the more complex work relationships that the industrial revolution and factory production were already beginning to create. In the political organization of his parishes he anticipated later anarchists by rejecting such standard democratic procedures as voting, since he regarded the rule of the majority as a form of tyranny. He not only envisaged society moving to a practice of consensus after its liberation from government, but also hoped that such a liberation would come into being through education and peaceful discussion. His anarchism was evolutionary rather than revolutionary.

The distinction between evolution and revolution is important since, apart from variations in their proposals for the economic organization of society, the main differences between the anarchists who began to appear with Proudhon were in their views of the necessary strategies for achieving the aim they all held in common – the abolition of the state and all forms of government, and their replacement by voluntary and cooperative forms of administration.

Some, like Leo Tolstoy, Henry David Thoreau and the Dutch anarchist leader, Domela Nieuwenhuis, were pacifists, aiming to change society by the practice of civil disobedience. Mohandas K. Gandhi, who more than once termed himself an anarchist and who envisaged a

decentralized society of village communes, was perhaps the most important of their company.

Proudhon was nearer to the pacifists in his view of the tactics of social change than he was to the later leaders of organized European anarchism. Though he often spoke of revolution, he hoped that peaceful change might come about through the creation of workers' economic organizations. Proudhon's mutualism, as he called it, was a mixture of peasant individualism and cooperativism aimed at the reorganization of society on an egalitarian basis. He set out to shock his readers by declaring that 'property is theft', but by this he really meant the use of property to exploit the labour of others. 'Possession' – the right of an individual worker or group of workers to control the land or machines necessary for production – he regarded as necessary for liberty. In the book that may be his masterpiece, *The General Idea of the Revolution in the Nineteenth Century,* written in prison because of his criticisms of Napoleon III, he sketched out the picture of a society of independent peasants and artisans with their small farms and workshops, and of factories and utilities like railways run by associations of workers, linked together by a system of mutual credit based on productivity and administered by people's banks like that which he attempted to establish during the revolution of 1848. Instead of the centralized state, he suggested a federal system of autonomous local communities and industrial associations, bound by contract and mutual interest rather than by laws, with arbitration replacing courts of justice, workers' management replacing bureaucracy, and integrated education replacing academic education. Out of such a pattern, Proudhon believed, would emerge the natural social unity which he equated with anarchy and in comparison with which, he believed, the existing order would appear as 'nothing but chaos, serving as a basis for endless tyranny'.

Proudhon was the real founder of the organized anarchist movement. He laid down its theoretical foundations in a continental European context where Godwin was virtually unknown, so that Mikhail Bakunin, possibly the best-known and most influential of anarchists, once admitted: 'Proudhon is the master of us all.' Proudhon's followers, who called themselves mutualists, were active in the foundation of the International Working Men's Association, the so-called First International, which provided the first of many battlegrounds between the authoritarian socialism of the Marxists and the libertarian socialism of the anarchists.

In the early days of the International the struggle was between Marx and his followers and the disciples of Proudhon, who had died in 1864, the year the International was founded. Later the struggle took a new form, since Proudhon's disciples were replaced in opposing Marx by the followers of Bakunin, a Russian aristocrat turned conspirator, and the conflict between them eventually destroyed the organization. It was basically the conflict between Marx's idea of the workers seizing control of the state to carry out the revolution, and Bakunin's idea of the workers carrying out the revolution in order to destroy the state and all the other manifestations of political power.

Bakunin accepted Proudhon's federalism and the argument in favour of working-class direct action, which the latter had developed in his final posthumously published work, *De la capacité politique des classes ouvrières* (The political capability of the working classes). But he argued that the modified property rights (the rights of 'possession') which Proudhon contemplated for individual peasants and artisans were impractical, and instead he proposed that the means of production should be owned collectively (hence his followers were called 'collectivists'). However, he still held like Proudhon that each man should be remunerated only according to the amount of work he actually performed; in other words, though in a slightly different form, the wages system would continue.

The second important difference lay in views of revolutionary method. Proudhon believed that one could create within existing society the mutualist associations that would replace it, and for this reason he came to oppose violent revolutionary action which aimed at an abrupt transition. Bakunin did not believe that such a piecemeal method could work. As a romantic revolutionary, he argued that 'the passion for destruction is also a creative passion', and taught that a violent

uprising was the necessary prelude to the construction of a free and peaceful society.

The individualism and non-violence implicit in Proudhon's vision were thrust into the side currents of anarchism; Tolstoy, who had known Proudhon, largely incorporated them in his teachings of a radical Christian anarchism. But down to the destruction of anarchism as a mass movement at the end of the Spanish Civil War in 1939, Bakunin's stress on violence and on a collectivized economic system remained dominant among anarchists in most countries.

The tactics of violent action varied, though they tended to be conditioned by the doctrine of propaganda by deed, which emerged during the 1870s among the Italian anarchists and was particularly propagated by Errico Malatesta. Individual assassinations, largely justified by this doctrine, became numerous around the turn of the century; a President of France and a President of the United States were among the victims. There were anarchist-inspired mass insurrections in Spain and Italy and, during the Russian Civil War, in the Ukraine, where for several years the anarchist leader Nestor Makhno established libertarian institutions over a wide area and protected them by a numerous Insurrectionary Army.

There were also variations in the concepts of collectivism which the anarchists pursued, exemplified particularly in anarchist communism and anarcho-syndicalism.

Anarchist communism was mainly developed by Peter Kropotkin, a Russian prince and a distinguished geographer who abandoned his privileges for the revolutionary cause, though the idea may have been developed first by the French geographer Elisée Reclus. Kropotkin wrote a number of the seminal works of anarchism, including *Mutual Aid: A Factor in Evolution,* in which he traced the development of cooperation among animals and men, and *Fields, Factories and Workshops,* in which he argued for the decentralization of industry that he considered an essential accompaniment to a non-governmental society.

The work in which Kropotkin most developed the idea of anarchist communism was *La Conquête du pain* (The conquest of bread), a kind of non-fictional utopia sketching out the vision of a revolutionary society organized as a federation of free communist groups. Kropotkin moved beyond Bakunin's collectivism, which envisaged common ownership of the means of production, to a complete communism in terms of distribution, which meant that need rather than merit would be the reason why a man should receive the means of life. Kropotkin argued that any payment according to the value of the work was a variant on the wages system, and that the wages system condemned man to economic slavery by regulating his patterns of work. Just as Kropotkin's anarchism was based on the idea (developed in *Mutual Aid*) that man was naturally social, so his idea of free communism was based on the notion that man was naturally responsible, and in a free society would neither shirk on his work nor take more than he needed from the common store.

Anarcho-syndicalism arose out of the involvement of anarchist activists in the French trade union movement, which revived during the 1880s after the proscriptions of working-class organizations that followed the Paris Commune of 1870. Industrial militancy seemed to offer a broad field for the direct action which the anarchists already advocated, and the anarcho-syndicalists tended to oppose to the gradualist tendencies of orthodox unionists, who sought the best possible deal with existing society, the intent to change that society by proceeding directly to the assumption of industrial control by the workers. Thus their unions, while not neglecting to fight for better conditions, were ultimately revolutionary in their intent, and a philosophy of incessant struggle developed among them. This concept was adapted by writers like Georges Sorel, who in *Réflexions sur la violence* suggested that the important aspect of revolutionary syndicalism was the myth of struggle and the cult of violence, which he believed had a regenerating effect on society. However, the working-class anarcho-syndicalist spokesmen, like Fernand Pelloutier, Emile Pouget and Paul Delesalle, rejected Sorel's theories, and believed that relentless industrial struggle, by violent and peaceful means, culminating in general strikes, could in fact destroy the capitalist system and the state at the same time. When that happened, the

syndicates would be transformed from organs of struggle into the organizational bodies of the new society, taking over places of production and organizing transport and distribution. In this way they were developing Proudhon's concept of mutualist institutions evolving within the society they would eventually replace. Anarchist purists, notably Errico Malatesta, distrusted the anarcho-syndicalists, fearing that a trade union movement that controlled all industry might itself be corrupted by power.

For many years before World War I, the anarcho-syndicalists controlled the leading French trade union organization, the CGT (Confédération Générale du Travail); after the war it was taken over by the communists, who had gained added prestige among the workers through the success of the Russian Revolution.

Anarcho-syndicalism, however, spread from France to Spain, where it became a powerful working class movement. The anarchist federation of unions (Confederación Nacional del Trabajo) was the largest labour organization in Spain, at times reaching more than two million members. It was a model of anarchist decentralization, employing only one paid secretary in its federal office, the actual tasks of organization being carried out in their spare time by workers chosen by their fellows. The CNT was strong among the peasants of Andalusia as well as in the factories of Catalonia. The civil war in 1936–39 brought Spanish anarchism to its apogee, which was followed quickly by its downfall. The experience of decades of street fighting enabled anarchist workers in the eastern cities of Spain to defeat the generals in the early days of Franco's military uprising. Later they sent their militia columns to the various fronts. At the same time they tried to bring about their anarchist millenium behind the lines by expropriating the factories and the large estates. Reports suggest that many of the factories were well run by the workers and that the collectivization of the land induced the peasants to work with pride and devotion. But the experiments were too brief for valuable conclusions to be drawn from them, since the anarchists' hatred of authority made them as inefficient in creating armies as they seem to have been efficient in organizing collective work, and their experimental communes were suppressed at the time of Franco's victory.

The outcome of the Spanish civil war led to a general decline of anarchism during the 1940s and 1950s. However, in the generally radical atmosphere of the 1960s it underwent a revival; anarchist groups appeared once again in Europe and North America, the movement's history was written by scholars, and the works of the great anarchist theoreticians appeared again in print. Anarchism has not become again a mass movement of the kind that once flourished in Spain and to a lesser degree in France, Italy and briefly in the Ukraine. But it is a visible movement once more. Anarchist ideas of decentralization have spread widely and have merged with those of the environmental movement. It now survives more as an intellectual trend, encouraging a critical view of the institutions and practices of authority, than as a quasi-apocalyptic movement which envisaged the end of government as a possible and not distant goal.

See Also

- ▶ Bakunin, Mikhael Alexandrovitch (1814–1876)
- ▶ Godwin, William (1756–1836)
- ▶ Proudhon, Pierre Joseph (1809–1865)

References

Joll, J. 1964. *The anarchists*. London: Eyre & Spottiswoode.

Marshall, P.H. 1984. *William Godwin*. New Haven: Yale University Press.

Masters, A. 1974. *Bakunin, the father of anarchism*. New York: Saturday Review Press.

Read, H. 1954. *Anarchy and order: Essays in politics*. London: Faber & Faber.

Rocker, R. 1938. *Anarcho-syndicalism*. London: Secker & Warburg.

Woodcock, G. 1956. *Pierre-Joseph Proudhon: A biography*. London: Routledge & Kegan Paul.

Woodcock, G. 1962. *Anarchism: A history of libertarian ideas and movements*. Cleveland: Meridian Books.

Woodcock, G., and I. Avakumovic. 1950. *The anarchist prince: A biographical study of Peter Kropotkin*. London: T.V. Boardman & Co.

Ancient Greece, The Economy of

Paul Cartledge

Abstract

There were many 'economies' rather than a single 'economy' in ancient Greece (a culturally interlinked world, c. 800–300 BCE, stretching across the Mediterranean basin and around the Black Sea). Except in Athens, agriculture (cereals, olives, grapevines, and the raising of small-stock animals – sheep, goats, pigs) predominated over trade and industry as an economic driver. The Greeks did not invent coinage but spread it and embedded it, and although they were thoroughly familiar with the idea of markets and market prices, they did not develop a market economy.

Keywords

Ancient economy; Oikos; 'Primitivists', 'modernists'; Finley, Moses; Athens; Sparta; 'Proxy data'; Mediterranean triad; Agriculture; Grain; Olive oil; Wine; Trade, local, regional and inter-regional; Manufacture; Technology; Slavery; Money, coined and non-coin; Markets

JEL Classifications

B11

Definitions

'Ancient Greece' for the purposes of this entry will be taken to refer to the period from roughly the eighth century BCE to the end of the fourth century BCE: that is, from the rediscovery of literacy by means of the invention of the Greek alphabet, the renewal of intensive trade contacts with the near East, and the beginnings of large-scale permanent overseas emigration and establishment of Mediterranean-wide trade networks, down to the start of the new post-Alexander the Great (d. 323 BCE) 'Hellenistic' (mixed Greek-oriental) world.

'Economy' is more difficult to specify, or pin down. The word is of Greek derivation but in ancient Greek it meant primarily and literally the management of an individual household (*oikos*) not the management of a 'city' or 'national' economy. This led one school of modern interpreters (late 19th-century German), the so-called 'primitivists', to speak of the entire period under consideration here as one of 'household' not 'national' economy. That is a considerable and highly misleading exaggeration, but it does draw attention to the fact that an ancient Greek city did not have an economy, or practise economics, in anything like a post-Adam Smith, let alone post-Alfred Marshall, sense. Their opponents, the 'modernists' or 'modernizers', claimed no less excessively that ancient Greece was, economically speaking, pretty much similar to the modern 'developed' world (and similar too in its stages of development), except that it operated on an infinitely smaller scale and without the benefits – such as economies of scale – made possible only by the scientific and technological revolutions of early modernity. A sensible compromise was firmly advocated by Moses Finley (1973), who rightly drew attention especially to Greek ideology and terminology; but because he sometimes underestimated the quantity and sophistication of ancient Greek economic activity, he too was (mis) labelled a 'primitivist'. (The debate is usefully summarized in Scheidel and von Reden 2002; see esp. Andreau 2002; Cartledge 2002; Meikle 2002; also Manning and Morris 2005.)

A second reason for being chary of the term 'economy' is that, after the wave of emigration noted above, there were at any one time between 600 and 300 BCE some 1,000 separate Greek political entities, radically self-differentiated politically but also often very different indeed economically speaking. This is why I have in the past written of 'the economy (economies) of ancient Greece' (Cartledge 2002; cf. Davies 1998). At one extreme,

classical fifth- and fourth-century Athens was as 'developed' as any Greek city before Hellenistic Alexandria (founded 332). In the international port of Peiraieus it even had a 'commercial centre' separated physically as well as spiritually from the political centre (Garland 2001), and its total population (civic centre plus surrounding territory) was far larger (c. 250–300,000) and far more diverse ethnically and occupationally than any other Greek city's. (In size of home territory, however, its c. 2,500 sq km were exceeded by Sparta, c. 8,400, Syracuse, c. 4,000, and Panticapaeum in the Crimea, c. 3,000: Hansen and Nielsen 2004, pp. 70–3). At the opposite extreme were fundamentally rural settlements, in Arcadia for example, cut off from the sea and long-distance trade, surviving, modestly, on 'natural', pastoral as well as agricultural, economy. In between, the modal Greek city had a population of some 2,000–8,000, occupied some 100 sq km, and practised versions of 'mixed economy', in which agriculture and stock-raising always predominated over trade and manufacturing industry.

A third problem with doing ancient Greek economic history is that the contemporary ancient data accessible today are resolutely unstatistical. This is partly because the ancient Greeks did not think and so did not audit themselves statistically but also because the nature of their politically overdetermined economies did not generally encourage or lend itself to statistical computation. The figures we get in our sources – for instance, for the impossibly inflated aggregate numbers of slaves in a particular city at any one time – tend therefore to be at best extreme outliers, at worst rhetorical inventions, rarely something reliably identifiable and usable in-between. Hence the regular resort of scholars to 'proxy data' – modern data of climate, crop-yields, etc. – making assumptions of continuity and stability as between ancient and modern conditions that are often untestable, but still useful as models or thought-experiments and for setting workable parameters (Manning and Morris 2005; on this and on all other matters discussed in this entry, see now Scheidel et al. 2007).

The Mediterranean Triad: Agriculture and Trade

The Mediterranean triad of dietary staples – grain, olive oil and wine – was established as such in the Greek sphere during the third millennium BCE (Renfrew 1972). Not much in the way of improvements in seed-selection, or efficiency in growing or harvesting techniques, is detectable during our period, owing to the likely constraints of Greek soils and microclimates, and the certain constraints of technological backwardness (not even the wheelbarrow was known, apparently).

One huge exception was the massively profitable exploitation after 600 BCE of the black-earth soils of the Ukraine and Crimea (see Panticapaeum, above) for the achieving of – by old Greek standards – huge yields of bread wheat, more nutritious as well as more easily processed than the default Greek grain-crop, barley (up to five times more drought-resistant than wheat on average) (Sallares 1991). Since the northern shore of the Black Sea cannot grow olives (because of winter frost), there was a considerable uplift in the production of olive oil further south (especially around Athens) for export to these deprived colonial Greeks, in exchange for which came, besides the bread wheat (especially again to Athens: Moreno 2007), dried fish and slaves.

Olives and their by-products were culturally as well as economically vital, and universally employed – even if not universally manufactured – as unguent, medicament, lubricant, and source of energy as well as food (Foxhall 2007). This was a standing incentive to extensification. In one small area near Athens, for example, terracing of marginal land is estimated to have extended the cultivated area by some 40 per cent. Regions and cities that experienced sharp population growth, such as Athens in the fifth century, also resorted to intensification, either by reducing the regularity or duration of fallowing, or by intercultivation of grain with olives, or by a combination of the two.

The grapevine flourishes in some soils and aspects more than others. The wine produced around Athens, for example, in sharp contrast to

its local olive oil, was thought far less desirable than that produced on the northern Aegean island of Thasos, where legislation was introduced in the fifth century to control the highly lucrative export trade. Other islands that specialized in wine-production for export were Lesbos, Chios and Samos, and each production region generated its own distinctive shape of two-handled, pointed-base, pottery transport vessels (known as *amphorae*), further distinguished by the liberal application of amphora-stamps – a sort of ancient Greek equivalent of *appellation contrôlée*.

Long-distance, interregional trade in wine, grain and other commodities (especially metals, such as copper from Cyprus or iron from Elba) was sharply marked off institutionally and terminologically from smaller-scale, local wholesale and retail trading (Garnsey et al. 1983). Long-distance traders were *emporoi*, literally 'passengers' (on ships), and they traded typically in purpose-built, 'round', sail-driven merchant vessels to economize on crew and time. But such trade was in Mediterranean weather conditions always risky, not to mention the threat from oar-driven pirate pinnaces and galleys; and from the later fifth century the larger operators were encouraged to take out bottomry or maritime loans – high-risk, high-interest – as a form of insurance, in deals struck with a new breed of commercially minded bankers (Cohen 1994). The owners of the banks as of the ships would be free, but the bankers and traders might just as likely be slaves as free citizens, and not rarely of non-Greek origin (Reed 2003).

Labour and Manufacture

The ancient Greek world was almost entirely one of human labour, not labour-saving technological devices, and, insofar as production was undertaken beyond the scale of household subsistence (Mattingly and Salmon 2001), it was a world of manufactories rather than factories. Wind power was of course exploited in navigation, but watermills were a thing of the pretty distant future so far as ancient Greece was concerned. Lifting devices and other forms of 'engineering' were most assiduously developed for religious not secular purposes, such as the construction of a temple (Landels 1978). On the other hand, traditional craft skills in carpentry, metallurgy, ceramics and the weaving of cloth had operated at a high level since the eighth century, even though typically on an individual household or small workshop basis (Burford 1972). One large exception were the gangs of workers employed in the silver-bearing lead mines belonging to the city of Athens, where possibly as many as 20,000 or even 30,000 may have been employed at any one time in digging, extracting and washing the ore (Lauffer 1979). But these were not free citizens: they were slaves, performing a task classified as servile, fit only for less-than-human beings.

Slavery

Unfreedom is of hoary antiquity, cross-culturally, but it took the ingenuity of the Greeks of the sixth century BCE to transform various kinds of personal dependency into full-blown ownership of the 'chattel' slave variety (the kind practised in the American Old South, the Caribbean and Brazil from the 17th to 19th centuries) (Dal Largo and Katsari 2007). At Athens, for example, all slaves were of the chattel variety, bought on the markets to which they had been brought by traders dealing with the countries of the Black Sea and western Anatolian regions (Scythians and Paphlagonians, for example). But in Sparta, although the same word (*douloi*) was used to describe them, the servile workforce of Helots ('captives') was created by enslaving local Greeks and by fair means or foul keeping them locked within a system of hereditary bondage for some four centuries or more (Luraghi and Alcock 2003). Almost all the 50–100,000 Helots, like the majority of the 100,000 or so slaves at Athens, were somehow engaged in agriculture. Just how economically efficient that system of helotage was cannot be easily determined, but it delivered the goods in the sense that it was the basis of Sparta's status as a great Greek power for most of those 400 years, and the basis too of Sparta's extraordinary warrior-communalist lifestyle.

Money and Coinage

Money has a number of functions and uses, some of which, but not all, can be handily fulfilled by coins (Howgego 1995). Greeks did not invent the idea of stamping a fixed weight of precious metal (gold, silver, electrum) with a badge and slogan or some other authenticating device, but they did develop this Lydian invention phenomenally, for political as well as economic reasons, and did transmit it to otherwise alien neighbouring cultures such as that of the Persians. The first coins struck, in the first quarter of the sixth century, were of electrum (a gold-silver mix) or silver, and of relatively large denominations, not usable therefore as small change. But by the end of that century sometimes really small fractions were in quite general use, at least in the Aegean, and by the end of the fifth century a move had been made towards a fiduciary coinage of bronze. The Spartans, idiosyncratic as ever, refused to strike silver coins (until their third-century BCE 'normalization'), but did operate some sort of 'currency' of iron (in the form perhaps of cooking spits). Such monetized spits were used elsewhere than in Sparta too, and offered as dedications to the gods and goddesses in sanctuaries, a nice reminder that the sacred and the profane were close partners in ancient Greece.

Markets

Finally, markets – and the issue of whether, and if so how far, any Greek city developed anything like a market economy: that is, not an economy with markets but an economy centrally defined by price-fixing markets. A famous passage of Aristotle's *Nicomachean Ethics*, written in the 330 s, has been read as making the intellectual breakthrough in embryo to a labour theory of value and/or a market theory of price, but it can just as well be read as an exercise in moral philosophy using economic illustrations. Elsewhere, in the *Politics*, Aristotle makes his preference for a 'free' Agora unambiguously plain: by 'free' he means one where sordid economic transactions were kept to a minimum, or at bay altogether – for example, by barring from the holding of political office any citizen who had traded in a commercial Agora within the past decade, as at Thebes (Austin and Vidal-Naquet 1977).

There is also objective evidence for a certain conventionality of non-market price-fixing – some commodities turn up in widely different contexts and periods valued at a suspiciously identical exchange price. Likewise, the cost of labour purchased on the market, for instance for large-scale civic construction projects such as temples, seems inelastic to a degree that would be considered irrational in an (economically speaking) free labour-market situation.

Nevertheless there are hints and signs from earlier in the fourth century of an increasing and increasingly generalized marketization of commodity exchange, a process that 'took off' exponentially under the conditions of the new world opened up by the middle Eastern conquests of Alexander the Great (334–323). But 'Hellenistic' economic globalization (as it were) is another topic, for another essay (see Cartledge 1997 for an overall outline sketch; in detail, Archibald et al. 2001, 2005).

Bibliography

Andreau, J. 2002. Twenty years after Moses I. Finley's *The Ancient Economy*. In Scheidel and von Reden, 2002.

Archibald, Z., J.K. Davies, V. Gabrielsen, and G.J. Oliver (eds.). 2001. *Hellenistic economies*. London/New York: Routledge.

Archibald, Z., J.K. Davies, and V. Gabrielsen (eds.). 2005. *Making, moving and managing: The new world of ancient economies, 323-31 B.C*. Oxford: Oxford University Press.

Austin, M., and P. Vidal-Naquet. 1977. *Economic and social history of ancient Greece: An introduction*. London: Batsford.

Burford, A. 1972. *Craftsmen in Greek and Roman society*. London/New York: Thames & Hudson.

Cartledge, P.A. 1997. Introduction. In *Hellenistic constructs: Essays in culture, history and historiography*, ed. P. Cartledge, P. Garnsey, and E. Gruen. Berkeley: University of California Press.

Cartledge, P.A. 2002. The economy (economies) of ancient Greece. In Scheidel and von Reden, 2002.

Cartledge, P.A., E.E. Cohen, and L. Foxhall (eds.). 2002. *Money, labour and land. Approaches to the economies of ancient Greece*. London/New York: Routledge.

Cohen, E.E. 1994. *The Athenian economy: A banking perspective*. Princeton: Princeton University Press.
Dal Largo, E., and C. Katsari (eds.). 2007. *Slave systems, ancient and modern*. Cambridge: Cambridge University Press.
Davies, J.K. 1998. Ancient economies: Muddles and models. In *Trade, traders and the ancient city*, ed. H. Parkins and C. Smith. London/New York: Routledge.
Finley, M.I. 1973. *The ancient economy*. Berkeley: University of California Press. (Latest edition by I. Morris, 1999).
Foxhall, L. 2007. *Olive cultivation in ancient Greece: Seeking the ancient economy*. Oxford: Oxford University Press.
Garland, R. 2001. *The Piraeus*, 2nd ed. Bristol: Bristol Classical Press.
Garnsey, P., K. Hopkins, and C.R. Whittaker (eds.). 1983. *Trade in the ancient economy*. London: Chatto & Windus.
Hansen, M.H., and T.H. Nielsen. 2004. *An inventory of archaic and classical Greek Poleis*. Oxford: Oxford University Press.
Howgego, C. 1995. *Ancient history from coins*. London/New York: Routledge.
Landels, J.G. 1978. *Engineering in the ancient world*. London: Chatto & Windus.
Lauffer, S. 1979. *Die Bergwerkssklaven von Laureion*, 2nd ed. Stuttgart: F. Steiner.
Luraghi, N., and S.E. Alcock (eds.). 2003. *Helots and their Masters in Laconia and Messenia: Histories, ideologies, structures*. Washington, DC: Center for Hellenic Studies.
Manning, J.G., and I. Morris (eds.). 2005. *The ancient economy: Evidence and models*. Stanford: Stanford University Press.
Mattingly, D.J., and J.B. Salmon (eds.). 2001. *Economies beyond agriculture in the classical world*. London/New York: Routledge.
Meikle, S. 2002. Modernism, economics and ancient history. In *The ancient economy*, ed. W. Scheidel and S. von Reden. New York: Routedge.
Moreno, A. 2007. *Feeding the democracy: The Athenian grain supply in the fifth and fourth centuries BC*. Oxford: Oxford University Press.
Parkins, H., and C.J. Smith (eds.). 1998. *Trade, traders, and the ancient city*. London/New York: Routledge.
Reed, C.M. 2003. *Maritime traders in the ancient Greek world*. Cambridge: Cambridge University Press.
Renfrew, C. 1972. *The emergence of civilization: The Aegean and the Cyclades in the third millennium B.C*. London: Methuen.
Sallares, R. 1991. *The ecology of the ancient Greek world*. London: Duckworth.
Scheidel, W., and S. von Reden (eds.). 2002. *The ancient economy*. Edinburgh: Edinburgh University Press.
Scheidel, W., I. Morris, and R. Saller (eds.). 2007. *The Cambridge economic history of the Greco-Roman world*. Cambridge: Cambridge University Press.

Anderson, James (1739–1808)

J. M. A. Gee

Keywords

Anderson, James; Corn Laws; Development economics; Edinburgh school; Poor Laws

Anderson farmed from the age of 15, first at Hermiston near Edinburgh, then at Monkshill, Aberdeenshire. Aberdeen honoured him with an LL.D. in 1780. He settled in Leith (near Edinburgh) in 1783, and founded *The Bee* (1790–94), a miscellany weekly magazine including literary, political and economic topics. He moved to London in 1797 and set up the magazine *Recreations* . . . (1799–1802) along the same lines as *The Bee*. The most important primary and secondary sources are listed below.

A contemporary of Adam Smith and James Steuart, James Anderson was second to none as a development economist. His writings lay great stress on the deadening effects of outmoded (feudal) institutions, adverse political and historic legacies, poor communications allied with sparse population, and repressive English-inspired taxation – especially the duties on salt and coal – on Scottish development. His proposals for improvement emphasized the gradualist approach – abstract economic models and grandiose schemes attracted his scorn – where the latent desire of man to improve his lot was freed from constraint and encouraged by state action and private self-interested philanthropy. Thus, though Anderson in general supported laissez-faire as being an essential requisite of optimal development, he believed the paternalistic encouragement of such development was frequently necessary, especially in the early stages. That he was no doctrinaire free-trader is seen in his espousal of the Corn Laws, on developmental grounds (see *An Inquiry into the Nature of the Corn Laws* . . .). He took issue with Smith on this, and also on Smith's notion that corn regulates the price of all commodities (see especially his

'Postscript to Letter Thirteen' in his *Observations* . . .). Smith never properly answered Anderson's criticisms (see Dow 1984).

Anderson is regarded as an anticipator of Ricardo's rent theory (see, e.g., Schumpeter 1954), but cannot in fact be cast in the narrowly abstract Ricardian mould. True, for Anderson an increase in corn price would have the differential effects on land rent as described by Ricardo; but this would be the first stage only of a development process. At the end of the process all land would have increased in fertility, and what was previously the least fertile cultivated land could well be now as fertile as the previously most fertile land (see *The Bee*, vol. 6, 28 December 1791).

Anderson was convinced of the harm caused by the Poor Laws, and was responsible for a successful appeal against the introduction of the poor rate in Leith.

In addition to his writings on agriculture and economic development and his literary magazine pieces, Anderson also wrote on slavery, archaeology and greenhouse and chimney design!

Selected Works

1777a. *Observations on the means of exciting a spirit of national industry.* . . . Edinburgh.
1777b. A*n inquiry into the nature of the Corn Laws.* . . . Edinburgh.
1785. *An account of the present State of the Hebrides, and Western Coasts of Scotland.* . . . Edinburgh.
1791–4. *The bee.* Edinburgh.
1794. *A general view of the agriculture and rural economy of the county of Aberdeen.* . . . Edinburgh.
1799–1802. *Recreations.* . . . London.

References

Anderson, William. 1865. *The Scottish nation*, vol. 1, 126–129. Edinburgh: A. Fullarton & Co.
Dow, A. 1984. The hauteur of Adam Smith: An unpublished letter from James Anderson of Monkshill. *Scottish Journal of Political Economy* 3: 284–285.
Mullet, C.F. 1968. A village Aristotle and the harmony of interests: James Anderson of Monks Hill. *Journal of British Studies* 8(1): 94–118.
Schumpeter, J.A. 1954. *History of economic analysis*. London: George Allen & Unwin.

Anderson, Oskar Nikolayevich (1887–1960)

Heinrich L. Strecker

Keywords
Anderson O. N.; Econometric Society; Econometrics; Index numbers; Quantity theory of money; Sample surveys; Statistics and economics; Tschuprow A. A.; Variate difference method

JEL Classifications
B31

Anderson was born on 2 August 1887 in Minsk, Russia, and died on 12 February 1960 in Munich, Federal Republic of Germany. As a disciple of Aleksandr A. Tschuprow the younger in St Petersburg, Anderson was a pioneer in statistics and econometrics. After leaving Russia in 1920 he became professor of statistics at the universities of Varna and Sofia in Bulgaria (until 1942), Kiel (until 1947) and Munich.

His oeuvre includes two textbooks and more than 150 articles in Russian, Bulgarian, English and German. Anderson participated during 1913–17 in the theoretical preparation and actual conduct of a sample on agricultural production in the Syr-Darja a river area of Russia, one of the very earliest sample surveys. Later, he designed the sample plan for the processing of the Bulgarian Agricultural Census of 1926, with very good results which were decisive for further propagation and acceptance of sampling (1929; 1949).

Before and after the First World War Anderson developed, independently of W.S. Gossett, the variate difference method, a procedure to separate

the smooth component (trend, business cycles) from the residual component, without making further assumptions about the underlying type of function (1929). Anderson wrote one of the first, much-noticed econometric papers, an effort to verify statistically the quantity theory of money, which was a very early analysis of causes by means of economic data (1931). Regarding index numbers, Anderson pointed particularly to the problem of chain index numbers, caused by error accumulation (1949; 1952).

Anderson was a charter member of the Econometric Society, a fellow or honorary member of numerous scientific associations, and held honorary doctorates from Vienna and Mannheim.

Selected Works

1929a. Über die repräsentative Methode und deren Anwendung auf die Aufarbeitung der Ergebnisse der bulgarischen landwirtschaftlichen Betriebszählung vom 31.
Dezember 1926. Munich: Fachausschuss für Stichprobenverfahren der Deutschen Statistischen Gesselschaft, 1949.
1929b. Die Korrelationsrechung in der Konjunkturfoschung. Ein Beitrag zur Analyse von Zeitreihen. *Bonn: Schroeder-Verlag.*
1931. Ist die Quantitätstheorie statistisch nachweisbar? *Zeitschrift für Nationalökonomie* 2, Vienna.
1935. *Einführung in die Mathematische Statistik*. Vienna: Springer-Verlag.
1949. Mehr Vorsicht mit Indexzahlen! *Allgemeines Statistisches Archiv* 33: 71–83, Munich.
1952. Wieder eine Indexverkettung? *Mitteilungsblatt für Mathematische Statistik* 4, Munich.
1954a. *Probleme der statistischen Methodenlehre,* 4th edn. Würzburg: Physica-Verlag. 1964.
1954b. Über den Umgang mit systematischen statistichen Fehlern. *Statistische Vierteljahresschrift* 7, Vienna.

Bibliography

Fels, E.M. 1968. Anderson, Oskar N. In *International encyclopedia of statistics,* vol. 1. London: Macmillan.
Strecker, H.L. 1965. Anderson, Oskar. In *Handwörterbuch der Sozialwissenschaften* (HDSW), vol. 12. Stuttgart/Tübingen: Fischer-Verlag/Mohr-Verlag.
Strecker, H., Kellerer, H., et al. 1963. *Oskar Anderson Ausgewählte Schriften,* 2 vols. Tübingen: Mohr-Verlag, with a complete bibliography of Anderson's publications.

Ando, Albert K. (1929–2002)

Charles Yuji Horioka

Keywords
Ando, A; Life-cycle hypothesis of saving; Modigliani, F; Simon, H

JEL Classifications
B31

Albert K. Ando was an eminent Japanese-born American economist who made many seminal contributions in a broad range of areas of economics. Born in Tokyo, Japan, on 15 November 1929, Ando went to the United States after the Second World War instead of joining the family business (ANDO Corporation, a major construction company). He received his BS in economics from the University of Seattle in 1951, his MA in economics from St Louis University in 1953, an MS in economics in 1956 and a Ph.D. in mathematical economics in 1959 from Carnegie Institute of Technology (now Carnegie Mellon University). After teaching at Carnegie and the Massachusetts Institute of Technology, Ando moved to the University of Pennsylvania in 1963 and remained there until his death from leukaemia on 19 September 2002, first as an associate professor of economics and finance, and from 1967 as a professor of economics and finance.

Ando held visiting appointments at universities in Louvain, Bonn and Stockholm, and consulted with the International Monetary Fund, the Federal Reserve Board, the Bank of Italy, and the Economic Planning Agency of Japan.

During his long and productive career, Ando received many honours and awards. For example, he was named Fellow of the Econometric Society, Ford Foundation Faculty Research Fellow, Guggenheim Fellow, and Japan Foundation Fellow, and was given the Alexander von Humboldt Award for Senior American Scientists.

Ando made important contributions in such diverse fields as econometrics (theory and applications), stochastic optimal control, the theory of aggregation and partitions in dynamic systems, monetary economics, macroeconomic modelling, and policy design, with an emphasis on interactions between economic growth and cyclical fluctuations, investment behaviour, theoretical and empirical investigations of household saving and consumption behaviour, and demography. His geographic breadth was equally great, with particular focus on Italy, Japan, and the United States. Ando collaborated, among others, with Nobel laureate Herbert Simon on questions regarding aggregation and causation in economic systems (see, for example, Simon and Ando 1961, and Ando et al. 1963) and with another Nobel laureate, Franco Modigliani, on extending the life-cycle hypothesis of saving (see, for example, Ando and Modigliani 1963), and constructing large-scale macroeconomic models (see, for example, Ando and Modigliani 1969).

A common thread in much of Ando's work is the care with which he analysed data. He subjected all of the data he used (whether national accounts data, data from household surveys, or company data) to careful scrutiny, was constantly on the lookout for inconsistencies, conceptual deficiencies, and so on, in the data, and made the necessary adjustments to the data to correct for any inconsistencies and conceptual deficiencies. He then analysed the resulting data meticulously and creatively to shed light on important questions such as the causes of the decade-long recession in Japan in the 1990s (he found that it was due primarily to the massive capital losses on household holdings of corporate equities; see, for example, Ando 2002a), whether aged households dissave (he found that they dissave relatively rapidly in Italy and the United States but moderately or not at all in Japan; see, for example, Ando and Kennickell 1987; Hayashi et al. 1988; and Ando and Nicoletti-Altimari 2004), how the cost of capital compares in the United States and Japan (he found that it is considerably higher in the United States if individual company data are used but not if national accounts data are used; see, for example, Ando and Auerbach (1988, 1990) and Ando et al. (1997).

Ando played a central role in the construction of the Massachusetts Institute of Technology, the University of Pennsylvania, and the Social Science Research Council (MPS) model, an early large-scale macroeconomic model of the US economy, as well of the Bank of Italy's macroeconomic model of the Italian economy (see, for example, Ando and Modigliani 1969, and Ando 1974), and in his later years he devoted considerable energy to constructing a dynamic microsimulation model of demographic structure for Italy, Japan and the United States, which he used to project future trends in the saving rate (he projected that Japan's saving rate would increase slightly in the immediate future as the number of children per family declined sharply, then fall moderately as the proportion of older persons in the population increased; he projected similar trends in Italy as well: see, for example, Ando et al. 1995, and Ando and Nicoletti-Altimari 2004).

See Also

- ▶ Modigliani, Franco (1918–2003)
- ▶ Simon, Herbert A. (1916–2001)

Selected Works

1961. (With H. Simon.) Aggregation of variables in dynamic systems. *Econometrica*. 29, 111–138.

1963. (With F. Modigliani.) The 'life cycle' hypothesis of saving: Aggregate implications and tests. *American Economic Review* 53, 55–84.

1963. (With F. Fisher, M. Franklin and H. Simon.) *Essays on social science models*. Cambridge, MA: MIT Press.

1965. (With F. Modigliani.) The relative stability of monetary velocity and the investment multiplier. *American Economic Review* 55, 693–728.
1965. (With G. Kaufman.) Bayesian analysis of the independent multinomial process – Neither mean nor precision known. *Journal of the American Statistical Association* 60, 347–58.
1969. (With F. Modigliani.) Econometric analysis of stabilization policies. *American Economic Review* 59, 296–314.
1974. Some aspects of stabilization policies, the monetarist controversy, and the MPS model. *International Economic Review* 15, 541–71.
1974. (With F. Modigliani, R. Rasche and S. Turnovsky.) On the role of expectations of price and technological change in an investment function. *International Economic Review* 15, 384–414.
1985. (With M. Blume, E. Marshall and I. Friend.) *The structure and reform of the US tax system*. Cambridge, MA: MIT Press.
1987. (With A. Kennickell.) How much or little life cycle is there in micro data? The cases of the United States and Japan. In *Macroeconomics and finance: essays in honor of Franco* Modigliani, ed. R. Dornbusch. Cambridge, MA/London: MIT Press.
1988. (With A. Auerbach.) The cost of capital in the United States and Japan: A comparison. *Journal of the Japanese and International Economies* 2, 134–158.
1988. (With F. Hayashi and R. Ferris.) Life cycle and bequest savings: A study of Japanese and US households based on data from the 1984 NSFIE and the 1983 Survey of Consumer Finances. *Journal of the Japanese and International Economies* 2, 450–491.
1990. (With A. Auerbach.) The cost of capital in Japan: Recent evidence and further results. *Journal of the Japanese and International Economies* 4, 323–350.
1992. (With L. Guiso, D. Terlizzese and D. Dorsainvil.) Saving among young households: Evidence from Japan and Italy. *Scandinavian Journal of Economics* 94, 233–250.
1995. (With A. Moro, J. Cordoba and G. Garland.) Dynamics of demographic development and its impact on personal saving: Case of Japan. *Ricerche Economiche* 49, 179–205.
1997. (With J. Hancock and G. Sawchuk.) Cost of capital for the United States, Japan and Canada: An attempt at measurement based on individual company records and aggregate national accounts data. In *Financing growth in Canada*, ed. P. Halpern. Calgary: University of Calgary Press.
2002a. Missing household saving and valuation of corporations: inquiry into Japanese national accounts I. *Journal of the Japanese and International Economies* 16, 147–176.
2002b. The elusive total budget outlay of the Japanese government: An inquiry into the Japanese national accounts II. *Journal of the Japanese and International Economies* 16, 177–193.
2004. (With S. Nicoletti-Altimari.) *A micro simulation model of demographic development and households' economic behavior in Italy.* Economic working paper No. 533. Economic Research Department, Bank of Italy.

Andreades, Andreas (1876–1935)

R. J. Bigg

Andreas (sometimes Andrew) Andreades was born in Corfu. His education (in France and England) and his academic affiliations were European, ranging from a doctorate in Law and Political Science from Paris University to the Bavarian Academy in Munich, the Romanian Academy in Bucharest and the Institut d'Egypte at Cairo. He became a lecturer at Athens University in 1902 and Professor of Economics in 1906.

The bulk of his writings were in Greek and French, effectively reducing his audience amongst English economists in the early 20th century. His interests were largely in the monetary and economic history of Greece. His financial history of ancient Greece was translated into English, but he was also concerned with contemporary Greek

problems and Eastern Europe. In the late 1920s and early 1930s he lectured widely in Europe (UK, France, Belgium, Italy) and in Egypt.

In England Andreades was perhaps best known for his *History of the Bank of England*, translated into English from the French in 1909. This was the first complete history of the Bank and Foxwell's introduction to the translation describes it as

> the best general survey of the subject which exists . . .. The author shows a remarkable familiarity with English methods and habits of thought, and his criticism is usually most just and temperate, and full of suggestion and stimulus (pp. xxiv–xxv).

Andreades attended the Paris and Danube Conferences in the interwar period and was a delegate to the assembly of the League of Nations in 1923, 1924 and 1929. He was chairman of the Greek League of Nations Union and president of the Athens branch of the Anglo-Hellenic League. He was honoured by the UK (CBE), Italy, Romania and Bulgaria, among other foreign countries.

Selected Works

1909. *History of the Bank of England*. Trans. C. Meredith, with Introduction by H.S. Foxwell. London: P.S. King & Son.

Andrews, Philip Walter Sawford (1914–1971)

Peter Earl

Andrews was born in Southampton and died in Lancaster. Most of his career was spent in Oxford and from 1946 until 1967, when he moved to his last post as Foundation Professor of Economics at the University of Lancaster, he was an Official Fellow of Nuffield College. He was founding editor of the *Journal of Industrial Economics*. In 1949, after conducting detailed case study investigations of business behaviour, Andrews published a potentially revolutionary analysis of firms in competitive oligopolistic markets. It included a non-marginalist, non-equilibrium theory of pricing and capacity choices. Firms were predicted to set prices by adding a mark-up to their 'normal' costs at their target levels of capacity utilization. The size of the mark-up would be limited by the difference between their own costs and their estimates of the opportunity costs of other firms with the knowledge to supply duplicates of their products and steal their markets. Only when their assessments of these cost conditions changed would they change their prices. Firms would also be expected to hold spare capacity in order to satisfy new customers without forcing established ones to turn their goodwill elsewhere.

After his death, Andrews's work was increasingly used as a building block in Post-Keynesian price theory. During his lifetime, however, his analysis failed to have a revolutionary impact, partly because most economists tried to make sense of it in orthodox terms; partly because Andrews generated confusion by writing in the language of business, not of textbook economics; and partly because it was not until 1964 that he published his incisive critique of the models he sought to displace.

Selected Works

1949. *Manufacturing business*. London: Macmillan.
1964. *On competition in economic theory*. London: Macmillan.

Angell, James Waterhouse (1898–1986)

Murray Milgate and Alastair Levy

From 1924 to 1966 Angell was a member of the faculty at Columbia University, but most of his

original work on monetary economics was undertaken in the decade between 1926 and 1936. This particular timing, together with the fact that Angell worked within the framework of the quantity theory of money, probably goes a long way towards explaining his comparative neglect in subsequent years – for this was the decade dominated by Keynes and his headlong assault on the quantity theory. Yet Angell was no mere expositor of that theory, and in his two most important books he contributed to its development in ways which were not to become fashionable until the influence of Keynesianism began to subside in the 1960s and 1970s.

Angell's first book, the *Theory of International Prices* (1926), was intended to provide a re-evaluation of classical theory in the light of the actual experience of the 18th and 19th centuries. Of the three main modifications he suggests – grounding the doctrine of comparative advantage on comparative money costs and prices rather than upon differential labour values; replacing the specie-flow adjustment mechanism with one based on adjustments via the domestic money supply and the price level; and the inclusion of the analysis of currency speculation in the determination of exchange rates – the last two are perhaps the most interesting. Angell's firm adherence to the quantity theory led him to appreciate that under the fixed exchange rate regime of the interwar gold standard, adjustment to international equilibrium had to be secured by movements in the domestic price level. On the opposite side of the Atlantic at about the same time, Keynes had made the same claim but had to waste much of his time in the famous debate over the return to gold in Britain simply trying to explain the point to his opponents. Of course, Keynes favoured abandoning fixed exchange rates in favour of managing the domestic money supply through Bank Rate policy, but in terms of his understanding of the international adjustment mechanism implied by the quantity theory his position was essentially the same as that of Angell.

Angell's other major contribution comes in his *Behavior of Money* published in 1936. This is an empirical study of the monetary history of the United States between 1890 and the 1930s. In it Angell analysed the relationship between the volume of bank deposits, the stock of notes and coins, the velocity of circulation, the general level of prices, and the volume of industrial production. He concluded that movements in nominal national income were highly correlated with changes in the stock of circulating medium. The velocity of circulation showed, he claimed, relative stability. With the customary 'real' forces determining real GNP, not only did this provide, to Angell's satisfaction, striking confirmation of the quantity theory, but it led him to a novel policy proposal: a quantitative rule for the restriction of the rate of change in the money supply. For Angell, 'the most effective . . . procedure [for] induc[ing] a greater stability in national and individual money incomes . . . is to stabilize the quantity of money itself' (1936, p. 163). It is relatively easy to see how closely both Angell's approach (an empirical analysis of actual monetary experience) and his specific policy prescription, anticipate the later work of Friedman and Schwartz. However, appearing as it did in the same year as Keynes's *General Theory*, it is not difficult to understand its lack of effect at the time.

It is not without interest to note that in his *Investment and Business Cycles* (1941), Angell directly criticized Keynes's theory of investment and that, in addition, Angell wrote a text on the interwar German recovery, *The Recovery of Germany* (1929). In 1945–6 he served as US representative on the Allied Commission on Reparations.

See Also

► Quantity Theory of Money

Selected Works

1926. *The theory of international prices: Criticism and restatement*. Cambridge, MA: Harvard University Press.
1929. *The recovery of Germany*. New Haven: Yale University Press.

1936. *The behavior of money: Exploratory studies*. New York/London: McGraw-Hill.
1941. *Investment and business cycles*. New York/London: McGraw-Hill.

Animal Spirits

Roger E. A. Farmer

Abstract

The term 'animal spirits' was used by Keynes to refer to the idea that business cycles might be caused by crowd psychology. Recent work, in the aftermath of rational expectations, has focused on incorporating this idea into general equilibrium theory by exploiting the fact that dynamic general equilibrium models often contain a continuum of indeterminate equilibria. In stochastic models, production may differ across states of nature solely because of differences in the rational self-fulfilling beliefs of investors. This dependence of outcomes on beliefs provides a modern interpretation of the idea that the business cycle may be driven by animal spirits.

Keywords

Animal spirits; Autarky; Capital–labour substitution; Commodity space; Comparative statics; Complete and incomplete participation; Determinacy and indeterminacy of equilibria; Dynamic inefficiency; Dynamic stochastic general equilibrium (DSGE) models; Externalities in preferences; Extrinsic uncertainty; General equilibrium; Golden rule; Great Depression; Hume, D.; Infinite horizon models; Insurance contracts; Intrinsic uncertainty; Keynes, J. M.; Kydland, F.; Leisure; Lucas, R.; Markov processes; Overlapping generations models; Prescott, E.; Rational expectations; Real business cycles; Returns to scale; Self-fulfilling expectations; Stabilization policy; Sunspot equilibrium; Technology shocks; Thornton, H.; Total factor productivity; Variable capacity utilization

JEL Classifications

D8

The term 'animal spirits' is closely associated with John Maynard Keynes, who used it in his 1936 book, *The General Theory of Employment Interest and Money*, to capture the idea that aggregate economic activity might be driven in part by waves of optimism or pessimism (although Robin Mathews 1984, p. 212, points out that Keynes would have been aware of its use by David Hume 1739, pp. 60–1).

> Most, probably, of our decisions to do something positive, the full consequences of which will be drawn out over many days to come, can only be taken as the result of animal spirits – a spontaneous urge to action rather than inaction, and not as the outcome of a weighted average of quantitative benefits multiplied by quantitative probabilities. (Keynes 1936, pp. 161–2).

The idea that waves of spontaneous optimism might drive business cycles was not new to Keynes and can be traced at least as far back as Henry Thornton, who attributed a central role in his theory of credit to '. . . that confidence which subsists among commercial men in respect to their mercantile affairs . . .' (Thornton 1802, p. 75).

The Advent of Rational Expectations

The early writers, including Keynes, did not develop fully worked-out dynamic models in which expectations of agents are related to outcomes that are later realized. The development of complete artificial economies of this kind occurred first with the rational expectations revolution in the 1970s in which the static macroeconomic disequilibrium model of Keynes's *General Theory* was replaced by modern dynamic general equilibrium models rooted in Chapter "Macfie, Alec Lawrence" of Gerard Debreu's *Theory of Value* (1959). This development began with the work of Robert E. Lucas, Jr., and early examples

of rational expectations models include Lucas and Leonard Rapping (1969) and Lucas (1972, 1973). Lucas's 1972 and 1973 papers were attempts to understand the business cycle as a monetary phenomenon. Monetary models gave way to exclusively real models of the business cycle following the publication of influential papers by Fynn Kydland and Edward C. Prescott (1982) and John B. Long and Charles Plosser (1983), and modern macroeconomics theories, based on these early contributions, are referred to as 'dynamic stochastic general equilibrium (DSGE) models'.

Early DSGE models were restricted to examples in which there exists a finite number of agents (often only one) choosing consumption, investment and employment sequences in an economy with complete markets. Infinite horizon (IH) models of this kind have the same structure as the finite general equilibrium model studied by Kenneth Arrow and Gerard Debreu (1954) and Lionel McKenzie (1959), with the exception that the commodity space is infinite dimensional. Timothy Kehoe and David Levine (1985) showed that the competitive equilibria of IH exchange economies satisfy the first and second theorems of welfare economics; and from applying their methods to production economies it follows that that consumption, investment and employment sequences can be treated 'as if' they were chosen by a social planner maximizing a concave objective function subject to a set of linear constraints. Social planning problems have a unique solution in which all fluctuations in investment must occur as a direct consequence of fluctuations in the fundamentals of the economy; typically taken to consist of preferences, endowments and technologies. It follows that, if expectations are rational, there is no room in these economies for animal spirits to exert an independent influence on economic activity.

The Infinite Horizon Model Under Constant Returns to Scale

The modern use of DSGE models has followed two routes. One class of models, following the IH approach, assumes that all decisions are taken by a finite set of infinitely lived households each of which makes decisions for current and future family members. This class includes the real business cycle (RBC) model, currently dominant in the profession, which has a history dating back to Frank Ramsey (1928), David Cass (1965) and Tjalling Koopmans (1965).

In simple representations of the IH model, one assumes that a single representative agent allocates output, Y_t between consumption, C_t and next period's capital stock, K_{t+1}. Output is produced from capital, K_t and labour L_t using a constant returns to scale technology that is subject to a productivity shock which is modelled as a random variable A_t. The representative agent ranks alternative probability distributions over consumption and labour supply using an additively separable utility function. This problem can be represented as follows:

$$\max_{\{C_t, L_t, K_{t+1}\}} \sum_{t=1}^{\infty} \left(\frac{1}{1+\rho}\right)^{t-1} E_1[U(C_t, L_t)], \quad (1)$$

$$Y_t = A_t K_t^a L_t^b, \quad (2)$$

$$K_{t+1} = K_t(1-\delta) + Y_t - C_t, K_1 = \overline{K}_1. \quad (3)$$

Here, $\rho > 0$ is the agent's discount rate and $0 \leq \delta < 1$ represents depreciation. The parameters a and b represent the elasticities of capital and labour in production and the assumption of constant returns to scale implies that

$$a + b = 1. \quad (4)$$

$E_1[\cdot]$ is the expectations operator, and the interpretation of this problem is that the agent chooses sequences $\{C_t(A^t), L_t(A^t)K_{t+1}(A^t)\}_{t=1}^{\infty}$ where $A^t = \{A_1, A_2 \ldots A_t\}$ is the history of shocks from date 1 to date t. A_t is a random variable, generated by an autocorrelated stochastic process.

In standard IH models one assumes that $U(x, y)$ is increasing in x, decreasing in y, strictly concave and twice continuously differentiable, and under these assumptions the programming problem defined in Eq. (1) is concave and has a unique solution. Under the commonly assumed functional form,

$$U(C,L) = \log(C) - \frac{L^{1+\gamma}}{1+\gamma},$$

this solution is characterized by the first order conditions:

$$C_t L_t^{\gamma} = b\frac{Y_t}{L_t}, \tag{5}$$

$$\frac{1}{C_t} = E_t\left\{\frac{1}{(1+\rho)C_{t+1}}\left(1-\delta+a\frac{Y_{t+1}}{K_{t+1}}\right)\right\}, \tag{6}$$

$$\lim_{T\to\infty}\left(\frac{1}{1+\rho}\right)^{T} E_1\left[\frac{K_{T+1}}{C_T}\right] = 0. \tag{7}$$

For the real business cycle programme it is critical to assume that the production function is linearly homogenous and preferences are strictly concave, since these assumptions imply that the problem of the representative agent has a unique solution. More generally, if there are multiple agents one can write down the problem of a social planner who maximizes a social welfare function, defined as a weighted sum of individual utilities.

The OLG Model and How it Differs

In contrast to the IH model, in overlapping generations (OLG) economies one assumes that the set of agents is infinite and that each agent lives for a finite number of periods; this model was developed first in English by Paul Samuelson (1958), although Maurice Allais' book (1948), written in French, predates Samuelson's contribution.

In OLG models, unlike in the IH model with concave preferences and technologies, there may exist equilibria that are dynamically inefficient. In equilibria of this kind the economy has 'too much capital', and a benevolent social planner could improve social welfare for all generations by consuming part of the capital stock (thereby raising consumption for the current generation) and diverting future output from investment to consumption (thereby raising consumption for all future generations).

After the publication of Samuelson's article in 1958, a considerable literature developed discussing the source of dynamic inefficiency. The question was finally settled with the publication of Shell's (1971) paper, 'Notes on the Economics of Infinity'. Shell argued that both IH and OLG models are special cases of Debreu's (1959) formulation of general equilibrium. In both cases the commodity space is infinite dimensional. In the IH model the number of agents is finite; in the OLG model it is infinite. This apparently innocuous difference is the key to understanding why there may be inefficient equilibria in the OLG model since, in an inefficient equilibrium, no single agent can make a welfare-improving trade. In contrast, dynamic inefficiency in an IH economy would imply the existence of an agent with infinite wealth at equilibrium prices.

Both IH and OLG models have been used as vehicles to develop the idea that animal spirits may independently influence economic activity. Since the IH model with concave preferences and technologies leads to equilibria that are efficient, it was the OLG model that was first exploited to develop the modern version of the 'animal spirits hypothesis'. However, since the period length of the two-period OLG model is typically interpreted as 25 or 30 years, and since the average period of a business cycle is six to eight years, it was easy to dismiss the early work, based on the OLG structure, on the grounds that the equilibria that it led to were theoretical curiosities that are not relevant in the real world. This criticism was addressed by a second generation of animal spirits economies, in which the OLG model was replaced by an IH framework that relaxed the assumption that the technology is subject to constant returns to scale.

Animal Spirits, Sunspots and Incomplete Participation

In DSGE models the term 'animal spirits' (Azariadis 1981; Howitt and McAfee 1992; Farmer and Guo 1994) is used interchangeably with 'sunspots' (Cass and Shell, 1983), 'self-fulfilling prophecies' (Azariadis 1981; Farmer 1993) and most recently 'irrational exuberance' by Alan Greenspan (1996) at an after-dinner speech.

Jevons (1884) used the term 'sunspots' to refer to the literal possibility that astronomical events could influence the trade cycle through the intermediating effect of the weather on agriculture. In their 1983 article, Cass and Shell meant something different. They constructed a two-period general equilibrium model with complete markets in which some agents are unable to enter into insurance contracts. They referred to this restriction as 'incomplete participation' to distinguish it from a potentially more serious market breakdown in which some kinds of insurance contracts cannot be entered into *by anyone*. Cass and Shell distinguished between *intrinsic* uncertainty, which can influence fundamentals of the economy, and *extrinsic* uncertainty, under which the fundamentals are unchanged across alternative extrinsic events. They showed that the inability of a subset of agents to enter into insurance contracts is a sufficient departure from standard general equilibrium assumptions to permit the existence of equilibria in which allocations differ across states of the world in which all uncertainty is extrinsic. When this occurs, they said that *sunspots matter*.

In an economy with a complete set of insurance markets and risk-averse agents, all of whom can participate in these markets, sunspots cannot matter. Since agents are risk averse, they would prefer the mean of a random allocation to the allocation itself. But if all uncertainty is extrinsic then the mean allocation is feasible; hence a sunspot allocation cannot be an equilibrium of a complete markets economy with complete participation. Sunspot equilibria are Pareto-inefficient, but for a different reason from the dynamic inefficiency associated with over-accumulation of capital in deterministic OLG models. Sunspot inefficiency arises from the addition of unnecessary randomness to an economy in which agents prefer to avoid fluctuations in their consumption allocations.

Animal Spirits in an OLG Model

The first application of sunspots to a DSGE model is due to Azariadis (1981). He constructed a two-period overlapping generations model with no intrinsic uncertainty. This model possesses a unique steady state in which money has value. Under typical assumptions about preferences, the linearized dynamics of equilibrium price sequences in the neighbourhood of the steady state obey a functional equation of the form

$$p_t = \alpha E_t\left[p_{t+1}\right] + c, |\alpha| < 1. \tag{8}$$

Azariadis looked for equilibria that follow a two-state Markov process: that is, equilibria of the form

$$\begin{bmatrix} p_t(s_t = 1) \\ p_t(s_t = 2) \end{bmatrix} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} \begin{bmatrix} \alpha p^1_{t+1}(s_t = 1) \\ \alpha p^1_{t+1}(s_t = 2) \end{bmatrix} + \begin{bmatrix} c \\ c \end{bmatrix} \tag{9}$$

where $s_t \in \{1, 2\}$ is the state at date t and π_{ij} is the probability that $s_t = i$ conditional on $s_{t-1} = j$. For the linearized model, the fact that $|\alpha| < 1$ implies that the only equilibrium in this class is one for which

$$p(s_t = i) = \frac{c}{1 - \alpha}, i = 1, 2, \tag{10}$$

that is, the price is constant and independent of the non-fundamental uncertainty. But in the nonlinear model the equation that defines equilibrium price sequences takes the form

$$p_t(s_t) = E_t\left[g\left(p_{t+1}(s_{t+1})\middle| s_t\right)\right], \tag{11}$$

where the function $g(\cdot)$ depends on assumptions about the form of the utility function. The equation defining a two-state Markov equilibrium takes the more general form

$$\begin{bmatrix} p_t(s_t = 1) \\ p_t(s_t = 2) \end{bmatrix} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} \begin{bmatrix} g\left(p^1_{t+1}(s_t = 1)\right) \\ g\left(p^1_{t+1}(s_t = 2)\right) \end{bmatrix}. \tag{12}$$

In this case, Azariadis showed that, as long as consumption and leisure are not gross substitutes, it is possible to find positive numbers p_1, p_2 such that $p_1 \neq p_2$ and positive probabilities π_{11}, π_{12}, π_{21} and π_{22} such that

$$\begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} \begin{bmatrix} g(p_1) \\ g(p_2) \end{bmatrix}. \quad (13)$$

In other words, prices (and implicitly employment, consumption and GDP) in this economy fluctuate between two different levels based purely on the occurrence of self-fulfilling expectations or, in Keynes's terminology, 'animal spirits'. As with the Cass–Shell example of sunspots, however, the Azariadis example could easily be dismissed as a model of a real economy since it required the assumption that consumption and leisure are gross complements – an assumption that was widely believed to be implausible and inconsistent with other evidence. The challenge was to develop a quantitative model of the business cycle in which aggregate fluctuations are driven by animal spirits, expectations are rational, and the model can capture the observed volatilities of output, consumption, GDP and hours.

Animal Spirits and Indeterminacy

The example of sunspots provided in the Cass–Shell (1982) paper relied on constructing an economy in which there are multiple equilibria. They showed that, when some agents are unable to participate in the insurance markets that occur before they are born, randomizations across deterministic allocations can also be sustained as equilibria. In the presence of complete participation in insurance markets these randomized equilibria would be ruled out since they are associated with unnecessary uncertainty that risk-averse agents would prefer to avoid.

In addition to the fact that an OLG equilibrium can be dynamically inefficient, there is a second key way in which OLG and IH models differ. In the IH model the set of equilibria is generically finite whereas OLG economies can contain a continuum of equilibria. (Roughly speaking, 'generically finite' means that for almost all IH economies there is a finite number of equilibria, and 'almost all' means that this statement is true for an open dense set of parameters in a parameterized family of economies.) The fact that there is a finite number of equilibria implies that each equilibrium of the IH model is locally unique, that is, there is no other equilibrium that is arbitrarily close to it.

A locally unique equilibrium is also called 'determinate'. Determinacy of equilibrium is an important property since, if one is interested in comparative statics, it is important that small changes in exogenous variables lead to predictable small changes in endogenous variables. If the equilibrium is one of a continuous set of equilibria (as would happen if the equilibrium were indeterminate) then the model does not make a clear prediction as to how prices and quantities would be expected to change in response to a change in policy or in some other fundamental of the economy.

Under some assumptions about preferences (a sufficient condition is that the endowment of the agents is sufficiently tilted towards youth), the one-good two-period OLG model possesses two steady states. Each of these steady states is a stationary equilibrium with a constant real rate of interest; in one stationary equilibrium money has positive value and in the other it does not. David Gale (1973) refers to economies that possess a monetary steady state as 'Samuelson' to distinguish them from those that do not (he calls these 'Classical'). In a Samuelson economy the two steady states are respectively 'generationally autarkic' (money has no value) and 'golden rule' (the real rate of interest equals the population growth rate). In Samuelson economies there exists a continuum of non-stationary equilibria and, when consumptions in adjacent periods are gross substitutes, each of these non-stationary equilibria converges to the autarkic steady state.

The non-stationary equilibria in the OLG model provide a rich source of equilibria over which to randomize; however, they all converge to an autarkic equilibrium in which money has no value. This property makes it difficult to construct stationary stochastic equilibria around the autarkic steady state since there are no non-stationary paths that approach the steady state from below. To get around this difficulty, Farmer and Woodford (1984) showed that, by adding government spending to the OLG model, one can construct randomizations over a set of non-stationary

equilibria that converge to a stationary state in which money has value. The addition of positive inflation-financed government expenditure shifts the set of stationary equilibria, and the indeterminate non-monetary equilibrium of the OLG model becomes a second monetary equilibrium. By adding a zero mean random variable to the model, Farmer and Woodford were able to construct a new set of *stationary* sunspot equilibria. Locally, these equilibria obey a difference equation of the form of Eq. (8), but the parameter α is greater than 1 in absolute value. It follows that one can construct equilibria in this model of the form:

$$p_{t+1} = \frac{1}{\alpha} p_t - \frac{c}{\alpha} + u_{t+1}, \tag{14}$$

where u_{t+1} is *any* random variable with zero conditional mean. Further, the unconditional probability distribution of the price level can be shown to converge to an invariant probability measure that depends on the distribution of the sequence of sunspot shocks, $\{u_t\}$. This is an important property of a rational expectations equilibrium since, arguably, stationarity is necessary for agents to learn about the world in which they live and to find ways of making unbiased forecasts of the moments of future prices.

Real Business Cycles and the Animal Spirits Hypothesis

The examples of stationary sunspot rational expectations equilibria, originally constructed in the OLG model, did not have much impact on mainstream macroeconomics. Although the first rational expectations models were constructed as monetary examples within the two-period OLG structure (for example, Lucas's seminal 1972 paper), the profession soon moved on to real models based on IH economies. The IH structure is more amenable to confrontation with data since the period of the model can easily be mapped into the period of data collection. Further, the examples of Azariadis and Farmer–Woodford were constructed in models that relied on assumptions widely believed to be unrealistic; these included the assumption of gross complements and two-period lives (in the case of the Azariadis model) and the assumption that sunspots exist close to a dynamically inefficient steady state in the Farmer–Woodford model (this assumption can be shown to generate counter-intuitive responses of inflation to expansionary fiscal policy).

To confront these criticisms, Howitt and McAfee (1992), Benhabib and Farmer (1994) and Farmer and Guo (1994) constructed examples of animal spirits equilibria within the IH paradigm by dropping the assumption that the technology is subject to constant returns to scale. At the time that this work was published, a number of authors (Caballero and Lyons 1993, are prominent examples) had estimated the degree of increasing returns to scale in US manufacturing industries and found it to be large.

In their 1994 paper, Benhabib and Farmer took a relatively standard IH model and added externalities and increasing returns to scale. Farmer and Guo (1994) constructed a discrete time version of the Benhabib–Farmer model and showed that it can be used to generate business cycle fluctuations driven by animal spirits. They argued that the animal spirits-driven model is *more* successful than the real business cycle model at capturing the observed dynamics of output, employment, investment and consumption because it can replicate the hump-shaped response of output and investment to shocks that is observed in US data.

The Benhabib–Farmer–Guo (BFG) model has the same form as the IH model described in Eqs. 11, 12, 13, 14, 15, 16, and 17 but it distinguishes between the *private* technology and the *social* technology. BFG assume that the economy contains a large number of identical firms, each of which produces output using the production function

$$Y_t = A_t K_t^a L_t^b. \tag{15}$$

In BFG, the term A_t is not exogenous. Instead, it represents an input externality of the form

$$A_t = \overline{K}_t^{\alpha - a} \overline{L}_t^{\beta - b}, \tag{16}$$

where $\overline{K_t}$ and $\overline{L_t}$ represent the economy-wide average use of capital and labour. Replacing

(1.16) in (1.15) and imposing the assumption that the economy is in a symmetric equilibrium in which $\overline{K_t} = K_t$ and $\overline{L_t} = L_t$ leads to the *social technology*

$$Y_t = K_t^{\alpha} L_t^{\beta}. \tag{17}$$

BFG assumed that

$$\alpha + \beta > 1, a + b = 1, \tag{18}$$

which implies that there are increasing returns to scale in the social technology but constant returns to scale at the level of the individual firm. Since increasing returns enter the economy as an external effect, each firm maximizes a concave profit function, and the equilibrium of the competitive economy is well defined. BFG showed that equilibria in their IH economy with increasing returns are characterized by the following system of equations.

$$Y_t = A_t K_t^{\alpha} L_t^{\beta}, \tag{19}$$

$$K_{t+1} = K_t(1-\delta) + Y - C_t, \tag{20}$$

$$C_t L_t^{\gamma} = b\frac{Y_t}{L_t}, \tag{21}$$

$$\frac{1}{C_t} = E_t\left\{\frac{1}{(1+\rho)C_{t+1}}\left(1 - \delta + a\frac{Y_{t+1}}{K_{t+1}}\right)\right\}, \tag{22}$$

$$\lim_{T\to\infty}\left(\frac{1}{1+\rho}\right)^{T} E_1\left[\frac{K_{T+1}}{C_T}\right] = 0. \tag{23}$$

When $a = \alpha$ and $b = \beta$, this model collapses to the real business cycle version of the IH economy. But if $\alpha > a$, $\beta > b$ and $\alpha + \beta$ is greater than 1 and 'large enough', Benhabib and Farmer showed that the dynamics of the IH model change character, and the model contains a continuum of indeterminate equilibria, just as the OLG model does. Farmer and Guo calibrated the model to US data and, by choosing parameters that appeared consistent with contemporary estimates of returns to scale, they showed that the model exhibits business cycles driven by self-fulfilling waves of optimism and pessimism.

To provide a degree of discipline to the calibration exercise, real business cycle economists estimate the volatility of real productivity shocks by constructing an estimate of total factor productivity (TFP). This is an accurate measure of TFP under the maintained assumptions of competitive markets and constant returns to scale. Farmer–Guo provided discipline to their calibration exercise by constructing the measure of TFP that would be estimated from data generated by an animal spirits economy by an econometrician who assumed incorrectly that the technology was driven by technology shocks, and imposed the incorrect identifying assumption of constant returns to scale. They showed that this measure has very similar properties to that of the TFP estimates from US data.

Animal Spirits, Business Cycles and Welfare

Much recent business cycle research assumes that business cycles are driven by technology shocks; but we do not have a very good explanation of what these shocks represent. The BFG model represents a plausible alternative to the real business cycle model. It recaptures an old idea and recasts it in modern language.

Why should we care if shocks arise in the productivity of the technology or in the minds of entrepreneurs? The answer is connected to the efficiency question. If business cycles arise as the consequence of the optimal allocation of resources in the face of unavoidable fluctuations in the technology, then there is not much that government can or should do about them. But, if they arise as the consequence of avoidable fluctuations in the animal spirits of investors, then the fluctuations that result are avoidable and the allocations are Pareto-suboptimal. Animal spirit-driven business cycles provide a reason for countercyclical stabilization policy, and the cause of cycles is therefore an important question.

In 1996 Takashi Kamihigashi showed that the RBC economy (driven by TFP shocks) and the Benhabib–Farmer model (driven by animal spirits) are observationally equivalent when

estimated on aggregate data and that, if one uses aggregate evidence alone, constant returns to scale is an identifying assumption. The empirical literature since the publication of volume 63 of the *Journal of Economic Theory* in 1994 suggests that early estimates of the degree of returns to scale were overstated, and more recent estimates (for example, Basu and Fernald 1997) are more modest. This has led to renewed developments by theorists who have constructed modifications of the basic animal spirits model that are able to bring down the required degree of returns to scale to well within the tolerance of the best econometric estimates. Innovations to this literature include the construction of multisector models (Benhabib and Farmer 1996; Weder 1998; Benhabib et al. 2000; Harrison 2001), externalities in preferences (Farmer and Bennett 2000; Hintermaier 2003), capital–labour substitution (Grandmont et al. 1998), stabilization policy (Schmitt-Grohé and Uribe 1997; Guo and Lansing 1998; Lloyd Braga 2003), alternative explanations of the Great Depression (Harrison and Weder 2006) and variable capacity utilization (Wen 1998; Benhabib and Wen 2004). Benhabib and Farmer (1999) provide a survey of this literature and references to additional related papers.

Recent examples of animal spirits-driven models are able to explain a wide range of phenomena and, when supplemented by the assumption of variable capacity utilization, the animal-spirits explanation of business cycles outperforms the RBC model in most dimensions. Since the two models have very different policy conclusions, research that addresses the question of whether business cycles are driven by animal spirits is likely to remain a lively and important focus of research for some time to come.

See Also

- ▶ Keynes, John Maynard (1883–1946)
- ▶ Keynesian Revolution
- ▶ Keynesianism
- ▶ Overlapping Generations Model of General Equilibrium
- ▶ Rational Expectations
- ▶ Sunspot Equilibrium

Bibliography

Allais, M. 1948. *Economie et intérêt*. Paris: Imprimerie Nationale.

Arrow, K., and G. Debreu. 1954. Existence of a competitive equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Azariadis, C. 1981. Self-fulfilling prophecies. *Journal of Economic Theory* 25: 380–396.

Basu, S., and J. Fernald. 1997. Returns to scale in US production: Estimates and implications. *Journal of Political Economy* 105: 249–283.

Benhabib, J., and R. Farmer. 1994. Indeterminacy and increasing returns. *Journal of Economic Theory* 63: 19–46.

Benhabib, J., and R. Farmer. 1996. Indeterminacies and sector specific externalities. *Journal of Monetary Economics* 37: 421–444.

Benhabib, J., and R. Farmer. 1999. Indeterminacy and sunspots in macroeconomics. In *The handbook of macroeconomics*, ed. J. Taylor and M. Woodford. Amsterdam: North-Holland.

Benhabib, J., K. Nishimura, and Q. Meng. 2000. Indeterminacy under constant returns to scale in multisector economies. *Econometrica* 68: 1541–1548.

Benhabib, J., and Y. Wen. 2004. Indeterminacy, aggregate demand and the real business cycle. *Journal of Monetary Economics* 51: 503–530.

Caballero, R., and R. Lyons. 1993. The case for external economies: An overview. In *Political economy, growth and business cycles*, ed. A. Cukierman, Z. Hercowitz, and L. Leiderman. Cambridge, MA: MIT Press.

Cass, D. 1965. Optimum growth in an aggregative model of capital accumulation. *Review of Economic Studies* 32: 233–240.

Cass, D., and K. Shell. 1982. Do sunspots matter? *Journal of Political Economy* 91: 193–227.

Debreu, G. 1959. *The theory of value: An axiomatic analysis of economic equilibrium*. New York: Wiley.

Farmer, R. 1993. *The macroeconomics of self-fulfilling prophecies*. Cambridge, MA: MIT Press.

Farmer, R., and R. Bennett. 2000. Indeterminacy with non-separable utility. *Journal of Economic Theory* 93: 118–143.

Farmer, R., and J.-T. Guo. 1994. Real business cycles and the animal spirits hypothesis. *Journal of Economic Theory* 63: 42–73.

Farmer, R., and M. Woodford. 1984. Self-fulfilling prophecies and the business cycle. Working paper no. 84–12. Caress, University of Pennsylvania. Reproduced in *Macroeconomic Dynamics* 1(1997), 740–769.

Gale, D. 1973. Pure exchange equilibrium of dynamic economic models. *Journal of Economic Theory* 5: 12–36.

Grandmont, J.-M., P. Pintus, and R. de Vilder. 1998. Capital–labor substitution and nonlinear endogenous business cycles. *Journal of Economic Theory* 80: 14–59.
Greenspan, A. 1996. The challenge of central banking in a democratic society. Remarks at the Annual Dinner and Francis Boyer Lecture of The American Enterprise Institute for Public Policy Research, Washington, D.C, December 5, 1996. Online. Available at http://www.federalreserve.gov/BoardDocs/speeches/1996/19961205.htm. Accessed 7 June 2006.
Guo, J., and K. Lansing. 1998. Indeterminacy and stabilization policy. *Journal of Economic Theory* 82: 481–490.
Harrison, S. 2001. Indeterminacy in a model with sector-specific externalities. *Journal of Economic Dynamics and Control* 25: 747–764.
Harrison, S., and M. Weder. 2006. Did sunspot forces cause the Great Depression? *Journal of Monetary Economics* 53: 1327–1339.
Hintermaier, T. 2003. On the minimum degree of returns to scale in sunspot models of the business cycle. *Journal of Economic Theory* 110: 400–409.
Howitt, P., and P. McAfee. 1992. Animal spirits. *American Economic Review* 82: 493–507.
Hume, D., and L. Selby-Bigge. 1739. *A treatise of human nature*. Oxford: Clarendon Press.
Jevons, W. 1884. On the study of periodic commercial fluctuations. In *Investigations in currency and finance*, ed. H. Foxwell. London: Macmillan.
Kamahigashi, T. 1996. Real business cycles and sunspot fluctuations are observationally equivalent. *Journal of Monetary Economics* 37: 105–117.
Kehoe, T., and D. Levine. 1985. Comparative statics and perfect foresight in infinite horizon economies. *Econometrica* 53: 433–453.
Keynes, J.M. 1936. *The general theory of employment interest and money*. London: Macmillan.
Koopmans, T. 1965. On the concept of optimal economic growth. In *The econometric approach to development planning*. Amsterdam: North-Holland.
Kydland, F., and E. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.
Lloyd Braga, T. 2003. Endogenous business cycles and systematic stabilization policy. *International Economic Review* 44: 895–915.
Long, J. Jr., and C. Plosser. 1983. Real business cycles. *Journal of Political Economy* 91: 39–69.
Lucas, R. Jr. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.
Lucas, R. Jr. 1973. Some international evidence on output–inflation tradeoffs. *American Economic Review* 63: 326–334.
Lucas, R. Jr., and L. Rapping. 1969. Real wages, employment, and inflation. *Journal of Political Economy* 77: 721–754.
Mathews, R. 1984. Animal spirits. *The Proceedings of the British Academy* 70: 209–229.
McKenzie, L. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27: 51–71.
Ramsey, F. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–559.
Samuelson, P. 1958. An exact consumption–loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66: 467–482.
Schmitt-Grohé, S., and M. Uribe. 1997. Balanced-budget rules, distortionary taxes, and aggregate instability. *Journal of Political Economy* 105: 976–1000.
Shell, K. 1971. Notes on the economics of infinity. *Journal of Political Economy* 79: 1002–1011.
Thornton, H. 1802. *An enquiry into the nature and effects of the paper credit of Great Britain*, 1978. Fairfield: August Kelley.
Weder, M. 1998. Fickle consumers, durable goods and business cycles. *Journal of Economic Theory* 81: 37–57.
Wen, Yi. 1998. Capacity utilization under increasing returns to scale. *Journal of Economic Theory* 81: 7–36.

Anthropometric History

John Komlos

Abstract

Anthropometric history is the study of human size as an indicator of how well the human organism fared during childhood and adolescents in its socio-economic and epidemiological environment. The development of this field has opened up new windows on the ways in which economic processes affected the populations experiencing it, such as the hidden costs of industrialization and urbanization.

Keywords

Antebellum puzzle; Anthropometric history; Biology; Cliometrics; Gross national product; Height; Biological welfare; Human Development Index; Labour productivity; Longevity; An indicator of biological welfare; Nutrition; Population density; Sen, A.; Slaves; Weight

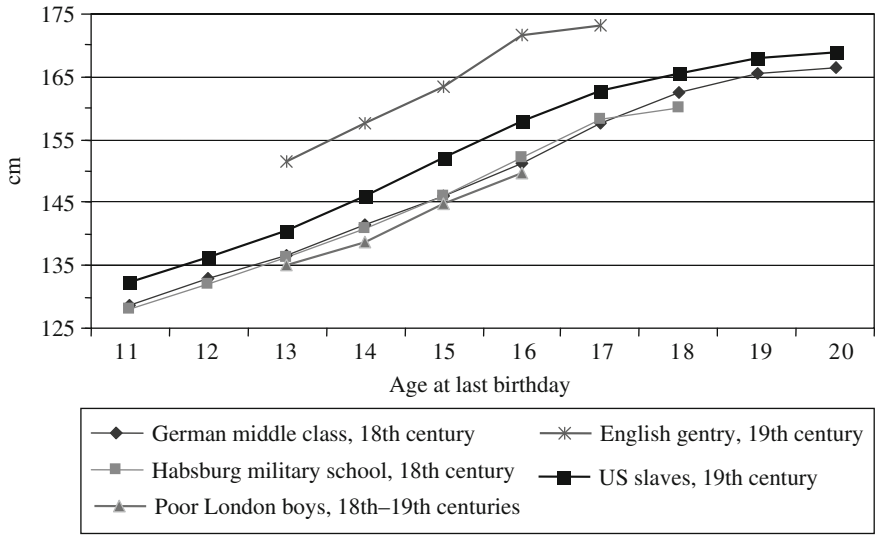
JEL Classifications

N0; I10; I31; N00; O10

Anthropometric history is the study of human size, primarily physical stature, weight, and the body mass index $\left[\frac{weight(kg)}{height(m)^2}\right]$ in order to ascertain how well the human organism thrived in its socio-economic and epidemiological environment.

As early as 1829 scholars recognized that the economy had a profound influence on human physical growth. In the 1960s French historians resurrected this tradition and explored the socio-economic correlates of height (Le Roy Ladurie et al. 1969), but the true expansion of the field began simultaneously in the mid-1970s among development economists and cliometricians. The former were interested in measuring malnutrition and its synergistic effect on economic performance in the Third World (Scrimshaw 2003). In cooperation with the United Nations, they expanded the work of nutritionists in combating poverty (Strauss and Thomas 1998) and measuring the impact of nutrition on labour productivity. Their effort culminated in the United Nation's formulation of the Human Development Index (HDI), which incorporates income, mortality, and schooling, in a superior measure of welfare (Sen 1987). In contrast, cliometricians analyse secular changes and cross-sectional patterns in biological welfare as well as the effect of economic development on the growth of the human organism. Initial research in this vein was influenced by the controversial finding that American slaves were relatively well nourished (Fogel and Engerman 1974), and was followed up by investigations of the height of slaves as an indicator of their nutritional status (Engerman 1976). The results implied that slaves were indeed well-nourished once they reached working age, as they were markedly taller than the European lower classes (Fig. 1) as well as their brethren in Africa (Steckel 1979). This astounding discovery prompted further research along these lines at a time when there was increased dissatisfaction with relying exclusively on gross national product (GNP) per capita as a welfare indicator, as it is not adjusted for income distribution or for externalities such as pollution; moreover, it pertains only indirectly to children and others not in the labour market, such as self-sufficient peasants and women for much of human history. Hence, GNP is only a rough indicator of well-being in a society.

The average height of a birth cohort – until adulthood is given approximately by:



Anthropometric History, Fig. 1 International comparison of height profiles (cm), 18th and 19th centuries (*Sources:* Steckel (1979), Komlos and Cuff (1998))

$$H(x)_t = H_{min}(x) + \int_{age=0}^{x} g\left[s_t \cdot Y_t, \left(\frac{P_f}{P_{aog}}\right)_t, W_t, D_t, \sigma_t, \theta_t, M_t, T_t, E_t\right] dt < H_{max}(x),$$

where $H(x)_t$ = physical stature at age = x for a particular birth cohort, for $x < 25$, Y_t = real disposable family income; s_t = share of income dedicated to children; P_f = price of nutrients; P_{aog} = price of all other goods (aog), W_t = work effort; D_t = epidemiological environment, σ_t = detrended variance of income longitudinally from $t = 0$ to $t = x$ (unpredictable income fluctuations might hinder the maintenance of an adequate diet). In turn, children sufficiently deprived will be forced off of their growth profile and may never catch up to their previous growth path; θ_t = cross-sectional inequality of income, M_t = cost of medical services, T_t = transfer payments from governments to families, E_t = environmental conditions (climate), and $H_{min}(x)$ and $H_{max}(x)$ are genetically determined minimum and maximum heights attainable by a given age; with

$$\frac{\partial g}{\partial Y} > 0, \frac{\partial^2 g}{\partial Y^2} < 0, \frac{\partial g}{\partial \frac{P_f}{P_{aog}}} < 0, \frac{\partial g}{\partial W} < 0, \frac{\partial g}{\partial D} < 0, \frac{\partial g}{\partial \sigma} < 0, \frac{\partial g}{\partial \theta} < 0, \frac{\partial g}{\partial M} < 0, \frac{\partial g}{\partial T} > 0$$

Diminishing returns to income imply that higher income volatility results in shorter stature for a given amount of average income over time. In practice, the analysis frequently pertains to the changes in height over time of adjacent cohorts of adults or of sub-adults of the same age in order to eliminate possible genetic components relevant to H_{min} (x) and H_{max} (x). Thereby one analyses how height is affected by the variables inside the integral over time (Komlos 1985; WHO 1995). Thus, adult height of a cohort reflects the history of its net-nutritional status during the growing years.

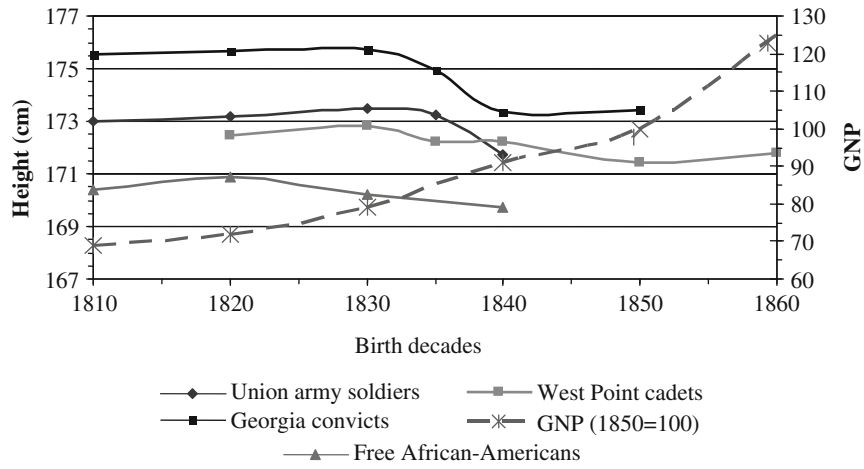
This innovative perspective opened up new windows to understanding of the impact of economic processes on the human organism and vice versa. According to archaeological evidence it is now evident that health of the natives of the New World '... was on a downward trajectory long before Columbus arrived' (Steckel and Rose 2002, p. 578). There were cycles in physical stature of about a generation long, brought about by demographic growth, urbanization, or changes in relative prices, market structure, income, inequality, and climate (Baten 2002; Baten and Murray 2000; Komlos 1998). There were also shorter cycles in height associated with business cycles (Woitek 2003); only in the 20th century were these cycles attenuated due to improvements in medicine, increases in labour productivity, and the substantial decline in the relative price of nutrients. The socio-economic crisis of the 17th century is evident in the height of the French population, as men measured about 162 cm on average (Komlos 2003). Europeans were never as short thereafter. The rapid population growth during the demographic revolution of the late 18th century brought about a decline in height everywhere in Europe as technological change in the agricultural sector did not suffice to maintain the nutritional status of the populations. The French Revolution was preceded by a decline in nutritional status, but no worse than in other parts of Europe, and not to the previous trough of the 17th century. Malthusian crisis generally began with a decline in heights even before mortality rates increased, as human organisms attempted to adjust their size to the available nutrition before the onset of subsistence crisis.

Social status has been related positively to height everywhere and at all times without exception. This generalization holds for 18th-century Germany as well as for the German Democratic Republic in the 20th century. The greatest social gradient in height ever recorded was found in early industrial England, where the difference between upper and lower class 15-year-olds reached 20 cm (Fig. 1). Height was related negatively to population density, as denser populations tended to have a higher disease load, as well as higher prices of nutrients. Urban populations tended to be shorter because of higher food prices and because of the higher incidence of diseases until the turn of the 20th century, when perishables became

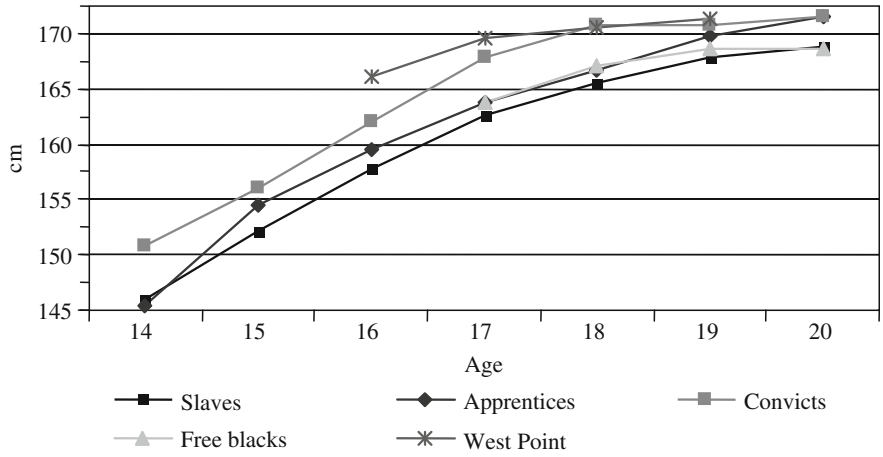
transportable longer distances due to refrigeration, and improvements in urban sanitation improved the epidemiological environment of towns. The degree of commercialization of the economy had an effect on human growth, as propinquity to nutrients invariably conferred considerable nutritional advantages in the early industrial period in so far as self-sufficient consumers did not have to pay for transportation costs of nutrients. Hence, self-sufficient (protein-producing) farmers tended to be tall. This was true in such widely separated places Tennessee, Japan or Bavaria (Cuff 1998; Craig and Weiss 1998; Haines 1998). Americans were the tallest in the world until the middle of the 20th century as resource abundance translated into higher wages, lower food prices, and a more equal distribution of income than prevailed elsewhere.

A transformation in the economic system put a hitherto unknown stress on the human organism. This was the case not only during the neolithic agricultural revolution but also during the Industrial Revolution, during the onset of modern economic growth as well as during the transition from socialism to capitalism. Thus, height declined (in the 1830s) at the onset of modern economic growth even in the resource-abundant United States, a phenomenon that has come to be known as the 'antebellum puzzle'. Average heights declined although real incomes increased (at a rate of 1.4 per cent per annum) because the relative price of nutrients and the degree of inequality were increasing and because self-sufficiency in agriculture was declining (Fig. 2).

Slaves were well nourished relative to the European lower classes (Fig. 1), even if they were not particularly tall in the US context (Fig. 3). Income was protective of nutritional status, as one would expect. High-status Americans did not experience a decline in height at the onset of modern economic growth, and the height of aristocrats did not decline during the Industrial Revolution. As Kuznets (1966) demonstrated, the anthropometric record also shows an increase in inequality with industrialization. Heights did not begin to improve substantially and reach their 18th-century levels until the end of the 19th century. Heights tended to correlate positively with wages except in the presence of countervailing forces. Height was associated positively with life expectancy up to about 185 cm; underweight and overweight individuals tended to have lower life expectancy; populations were underweight prior to the mid-20th century as food was relatively expensive and there was a lot of physical activity associated with daily life. Much of the increase in life expectancy in the 20th century is associated with an increase in body size; however, for the first time in its existence, because of technological and cultural



Anthropometric History, Fig. 2 The puzzling trend in the height of Americans during the antebellum period (*Sources*: Margo and Steckel (1983), Komlos (1987, 1998), Komlos and Coclanis (1997), Weiss (1994))



Anthropometric History, Fig. 3 Height of US youth, early 19th century (*Sources*: Komlos (1987, 1998), Komlos and Coclanis (1997), Steckel (1979))

changes the human species is facing an obesity epidemic that threatens to slow down the rate of increase of life expectancy.

The citizens of the western and northern European welfare states are the tallest in the world now, having overtaken the Americans about a generation ago. That implies that these welfare states provide a higher biological standard of living than the more free-market-oriented American society (Komlos and Baur 2004).

With the development of the concept of the 'biological standard of living' as distinct from conventional indicators of well-being, and with the founding of the new journal *Economics and Human Biology* in 2003, biology became integrated into mainstream economics. Height and weight are components and relatively easily measured indicators of biological welfare. In addition, we gain new insights of the effect of economic processes on the human organism. Hence, anthropometric history emphasizes that well-being encompasses more than the command over goods and services. Rather, it is multidimensional, and height, weight, health in general, and longevity all contribute to it – independently of purchasing power. In many ways, such indexes provide a more nuanced view of the impact of dynamic economic processes on the quality of life than income or GNP per capita alone. Anthropometric indicators are not meant to be substitutes for, but complements to, such conventional measures of living standards as income per capita.

See Also

- ▶ Cliometrics
- ▶ Development Economics
- ▶ Economic History
- ▶ Environmental Kuznets Curve
- ▶ Family Economics
- ▶ Fogel, Robert William (Born 1926)
- ▶ Industrial Revolution
- ▶ Nutrition and Development
- ▶ Sen, Amartya (Born 1933)

Bibliography

Baten, J. 2002. Climate, grain production and nutritional status in 18th century southern Germany. *Journal of European Economic History* 30: 9–47.

Baten, J., and J. Murray. 2000. Heights of men and women in nineteenth century Bavaria: Economic, nutritional, and disease influences. *Explorations in Economic History* 37: 351–369.

Craig, L., and T. Weiss. 1998. Nutritional status and agricultural surpluses in the antebellum United States. In *The biological standard of living in comparative perspective*, ed. J. Komlos and J. Baten. Stuttgart: Franz Steiner.

Cuff, T. 1998. Variation and trends in stature of Pennsylvanians, 1820–1860. In *The biological standard of*

living in comparative perspective, ed. J. Komlos and J. Baten, Vol. 1. Stuttgart: Franz Steiner.
Engerman, S. 1976. The height of US slaves. *Local Population Studies* 16: 45–49.
Fogel, R., and S. Engerman. 1974. *Time on the cross: The economics of American Negro alavery*. Vol. 2. Boston: Little, Brown and Co.
Haines, M. 1998. Health, height, nutrition, and mortality: Evidence on the 'antebellum puzzle' from Union army recruits for New York State and the United States. In *The biological standard of living in comparative perspective*, ed. J. Komlos and J. Baten. Stuttgart: Franz Steiner.
Komlos, J. 1985. Stature and nutrition in the Habsburg Monarchy: The standard of living and economic development in the eighteenth century. *American Historical Review* 90: 1149–1161.
Komlos, J. 1987. The height and weight of West Point Cadets: Dietary change in antebellum America. *Journal of Economic History* 47: 897–927.
Komlos, J. 1998. Shrinking in a growing economy: The mystery of physical stature during the Industrial Revolution. *Journal of Economic History* 58: 779–802.
Komlos, J., in collabaration with M. Hau and N. Bourguinat. 2003. An anthropometric history of early modern France, 1666–1766. *European Review of Economic History* 7: 159–189.
Komlos, J., and M. Baur. 2004. From the tallest to (one of) the fattest: The enigmatic fate of the size of the American population in the twentieth century. *Economics and Human Biology* 2: 57–74.
Komlos, J., and P. Coclanis. 1997. On the puzzling cycle in the biological standard of living: The case of the antebellum Georgia. *Explorations in Economic History* 34: 433–459.
Komlos, J., and T. Cuff. 1998. *Classics of anthropometric history*. St. Katharinen: Scripta Mercaturae.
Kuznets, S. 1966. *Modern economic growth: Rate, structure, and spread*. New Haven: Yale University Press.
Le Roy Ladurie, E., N. Bernageau, and Y. Pasquet. 1969. Le conscrit et l'ordinateur: Perspectives de recherches sur les archives militaires du XIXe siècle français. *Studi Storici* 10: 260–308.
Margo, R., and R. Steckel. 1983. Heights of native born Northern whites during the antebellum period. *Journal of Economic History* 43: 167–174.
Scrimshaw, N. 2003. Historical concepts of interactions, synergism and antagonism between nutrition and infection. *Journal of Nutrition* 133(1): 316S–321S.
Sen, A. 1987. *The standard of living*. Cambridge: Cambridge University Press.
Steckel, R. 1979. Slave height profiles from coastwise manifests. *Explorations in Economic History* 16: 363–380.
Steckel, R. 1995. Stature and the standard of living. *Journal of Economic Literature* 33: 1903–1940.
Steckel, R., and J. Rose, eds. 2002. *The backbone of history: Health and nutrition in the Western Hemisphere*. New York: Cambridge University Press.
Strauss, J., and D. Thomas. 1998. Health, nutrition, and economic development. *Journal of Economic Literature* 36: 766–817.
Weiss, T. 1994. Economic growth before 1860: Revised conjectures. In *Economic development in historical perspective*, ed. D. Schaefer and T. Weiss. Stanford: Stanford University Press.
WHO (World Health Organization). 1995. *Physical status: The use and interpretation of anthropometry*. Technical Report Series No. 854. Geneva: WHO.
Woitek, U. 2003. Height cycles in the 18th and 19th centuries. *Economics and Human Biology* 1: 243–258.

Anti-Discrimination Law

John J. Donohue III

Abstract

This article reviews the major legislative initiatives outlawing discrimination, discusses the theoretical arguments for and against such initiatives, and assesses the impact of these laws on the groups they try to protect. The significant effects of federal law in the first decade after passage of the 1964 Civil Rights Act are contrasted with the less optimistic findings from subsequent anti-discrimination interventions. Insights about the social benefits and the costs of the unintended consequences of employment discrimination law apply equally to other types of antidiscrimination legislation, such as mortgage lending and policing.

Keywords

Anti-discrimination law; Becker, G; Black–white labour market inequality in the United States; Efficient capital markets hypothesis; Friedman. M; Human capital; Insider–outsider problems; Jim Crow South; Labour market discrimination; Sex discrimination; Statistical discrimination

JEL Classifications

K3

In the aftermath of the Second World War, New York and New Jersey became the first in a series of non-Southern states to pass laws prohibiting racial discrimination in employment. Almost two decades later Congress passed, over strong Southern opposition, the momentous Civil Rights Act of 1964, which banned discrimination on the basis of race, sex, religion and national origin in employment and public accommodations. Over the ensuing 40 years, the reach of federal and state antidiscrimination law has extended beyond intentional discrimination (disparate-treatment discrimination) to ostensibly neutral practices that have an adverse impact on selected groups (disparate-impact law), and to protect those over age 40 (the Age Discrimination in Employment Act) and those with disabilities (the Americans with Disabilities Act). Anti-discrimination law has come to play an increasingly important role in employment, government contracting, policing and criminal justice, mortgage lending, retail and marketing practices, and education.

The Becker Model, Federal Anti-Discrimination Law, and the End of the Jim Crow Era

In 1957, Gary Becker launched the serious economic evaluation of discrimination when he developed a model based on individual animus towards a certain class of workers (see Becker 1957). The analysis had a number of shortcomings when applied to the real world, not least of which was that it assumed that the psychological burden of discrimination fell only on the *discriminators* (they were the ones who suffered the distaste), and the only cost borne by the victims of discrimination was any resulting decrease in wages or employment. In the early 1960s, Milton Friedman, in part influenced by Becker's work, argued against employment discrimination law on the grounds that it was unnecessary since competitive markets would protect workers from discrimination, and undesirable since government should not interfere with the personal preferences of discriminating employers. Although it is now clear that Friedman's position was incomplete, both arguments carry some weight.

First, frictionless competitive markets should offer protection from discriminatory employers. This means that, even in the presence of substantial employer animus, highly competitive markets reduce the need for law if a sufficient number of non-discriminators are available to bid up the wages of, say, black labour. The efficient capital markets hypothesis assumes that prices of financial assets will always tend to be close to their underlying value. Workers are also valuable assets, so Friedman believed that competitive markets would similarly push wages towards underlying productivity ('true value') in the labour market as well. But, even under the best of circumstances, one would not expect labour markets to be as efficient as capital markets with their homogeneous products, low transaction costs, ability to sell short and hordes of analysts whose job it is to identify the true value of certain securities. The resulting trades will tend to push these stock prices towards their true value (Donohue 1994). In the labour market, workers are not homogeneous, transaction costs associated with hiring and dealing with labour are high, there is no ability to sell short, and the value in ascertaining the true productivity of a modal worker is relatively small. If one adds in labour market imperfections posed by unions, minimum wage laws, high information costs and the racist and segregationist Jim Crow laws – laws requiring strict racial segregation in many aspects of public life including schooling and accommodations that led to inferior treatment of blacks despite the supposed legal requirement of equality under the 'separate but equal' doctrine – it is not hard to imagine that, in the absence of anti-discrimination legislation, blacks would be unfairly excluded from a range of good jobs or paid less than their marginal product.

Moreover, while competitive markets would be hostile to employer discrimination, they would actually encourage an employer to discriminate if that is the preference of fellow workers or customers. Moreover, the empirical evidence demonstrated clearly that, whether from the

pressures of racist norms or governmental encumbrances, the market afforded little protection to black workers in major industries of the South, such as Southern textiles (Heckman and Payner 1989). The major federal intervention directed against the Jim Crow policies of the South beginning with the 1964 Civil Rights Act did what competitive markets had failed to achieve – open up entire industries to qualified black workers and substantially dampen the black shortfall in earnings vis-à-vis white workers (Donohue and Heckman 1991).

Second, under the Becker model, net utility will be decreased by an employment discrimination law if one gives weight to the preferences of discriminators, as Friedman and Becker were wont to do. But Donohue (1986, 1989) argued that driving the discriminators out of business could actually enhance welfare by eliminating the Beckerian social cost. Moreover, while Becker conceived of discrimination as a stable taste, the evidence again suggests that the federal prohibition ultimately changed the attitudes (tastes) of millions of Americans. Rather than relentlessly and constantly imposing the burdens of inefficient interactions on unwilling discriminators, the Civil Rights Act aided a social process of integration that ultimately reduced the prior Beckerian taste for discrimination. While short-run costs were undoubtedly high, in the long run an entire region of the country was energized by the disruption of previously regimented views of racial inferiority – to the benefit of both blacks *and whites*. Since the Beckerian discriminatory tastes represented social costs, the reduction in the magnitude of these social costs constituted a major social benefit.

Did Federal Law Improve the Economic Status of Blacks and Others?

Perhaps the most important question concerning federal anti-discrimination law is whether it has aided its primary intended beneficiaries – black Americans (particularly in the South). James Smith and Finis Welch (1989) argued that the Civil Rights Act of 1964 was not responsible for substantial gains in black economic welfare. They conceded that black economic welfare improved at about the time of the federal initiatives in the 1960s, but they contended that the gains were the result of human capital enhancement, not of demand-side policies addressed to ameliorate the impact of discrimination. To buttress their view that Title VII – the section in the Civil Rights Act prohibiting employment discrimination based on, inter alia, race or colour – generated no benefits for black workers, Smith and Welch argued that the economic gains of blacks during the period 1940–60 were the same as those in the 1960–80 period (thereby suggesting that the Civil Rights Act of 1964 had been unimportant). The major response to Smith and Welch came from Donohue and Heckman (1991), who argued that Title VII did indeed generate a decade of economic gains for blacks:

> . . .the evidence of sustained economic advance for blacks over the period 1965–1975 is not inconsistent with the fact that the racial wage gap declined by similar amounts in the two decades following 1940 as in the two decades following 1960. The long-term picture from at least 1920–1990 has been one of black relative stagnation with the exception of two periods – that around World War II and that following the passage of the 1964 Civil Rights Act. (Donohue and Heckman 1991, p. 1614)

It is now widely accepted that, in helping to break down the extreme discriminatory patterns of the Jim Crow South, Title VII considerably increased the demand for black labour, leading to both greater levels of employment and higher wages in the decade after its adoption (see also, Freeman et al. 1973; Conroy 1994; and Orfield and Ashkinaze, 1991). Chay (1998) shows that, when the reach of the 1964 Civil Rights Act was expanded in 1972, the demand for black labour was further stimulated. But the good news in terms of law-induced efforts to improve the economic status of blacks through anti-discrimination policy has probably run its course. A series of papers by Oyer and Schaefer (2000, 2002a, b) offers little support for the view that the strengthening of federal anti-discrimination law in 1991 stimulated black or female employment, as occurred with the federal laws passed in 1964 and 1972. (The CRA actually changed race

discrimination law in a relatively minor way – restoring the standards that had existed in June 1989 with respect to discriminatory discharge and the standards for employer justification of practices with disparate racial impacts. For non-race cases, however, the 1991 Act expanded the damages available and authorized punitive damage awards for intentional discrimination.)

Moreover, papers by Acemoglu and Angrist (2001), and DeLiere (2000) hold that another piece of anti-discrimination legislation, the Americans with Disabilities Act (ADA), actually harms employment. This very pessimistic conclusion may be too strong. Attributing the poorer employment experience of the disabled in a short period after the federal law passed in 1990 turns out to be a tricky proposition, given the downturn in the economy and the substantial growth in those collecting disability benefits at roughly the same time. Burkhauser et al. (2006) extend the time period of Acemoglu and Angrist's analysis, and conclude that the decline in relative employment of the disabled actually began in the mid-1980s, roughly the time at which rules for disability benefits eligibility were loosened. But even if the ADA did not hurt, there is no strong evidence that it helped on the macro level, even if it did assist in securing small micro-level accommodations for the disabled. Jolls and Prescott (2004) argue that disability laws having a reasonable accommodation requirement may generate an insider–outsider problem. Those who gain the accommodation are better off, but at the expense of some disabled workers who end up out of the labour force.

Is Employment Discrimination a First-Order Problem for US Blacks Today?

Is the Black–White Earnings Differential Fully Explained?

Heckman (1998) contends that labour market discrimination no longer substantially contributes to the black–white wage gap (as it once clearly did), and therefore he doubts that four decades after the Civil Rights Act racial discrimination in the labour market is a first-order problem in the United States. Rather, Heckman looks to other factors (namely, those that promote skill formation) to explain the black–white earnings gap – a theme that he builds on in Carneiro et al. (2005).

An important paper that informs Heckman's analysis of the current reasons for the black–white wage gap is Neal and Johnson (1996). If factors that exist prior to workers' entry into the labour market largely explain the black–white wage gap, then the contribution of racial discrimination to this wage gap is presumably small. Neal and Johnson note that many studies have examined the black–white wage gap and found that it could not be explained with standard measures such as age, years of education, marital status and so forth, implying that the contribution of discrimination was sizable. Neal and Johnson note that years of education may exaggerate the true skill level attained by blacks, given the poorer-quality schools that many blacks attend. They argue that scores on the Armed Forces Qualification Test (AFQT) are a better measure than innate ability of acquired skill brought to the labour market.

The authors begin by showing that the *unadjusted* wage gap between blacks and whites is minus 24.4 per cent for black men and minus 8.5 per cent for black women. Using National Longitudinal Surveys of Youth (NLSY) data, Neal and Johnson found that the unexplained wage gap fell to minus 7.2 per cent for black men and plus 3.5 per cent (although insignificant) for black women, once they controlled for race, age and AFQT score. In other words, the AFQT test score can explain a very large portion of the black–white wage gap for men, and the entire gap for women. One source of continuing debate in the literature is whether these wage regressions should include controls for years of education as well as AFQT score. Neal and Johnson say it should not since the test better captures ability, and so they exclude the education measure from their regressions. Others have included years of education and find that the unexplained wage gap re-emerges when this control is added.

A potential problem with their approach is the possibility of black underinvestment in human capital due to the presence of statistical discrimination. Neal and Johnson reject this concern,

finding that the return to higher AFQT scores is significantly higher for black men (although not for black women), so that blacks seem to have adequate incentive to invest in developing human capital.

Evidence of Racial Discrimination in Entry Level Hiring from Audit-Pair Studies

The view that racial discrimination seems to have largely been wrung from the labour market is in apparent conflict with a number of audit studies that document differential treatment of blacks and whites. For example, a recent study by Devah Pager concludes that the degree of discrimination in employment is so great that blacks without criminal records are treated as badly as whites with criminal records (Pager 2003). Pager's audit experiment involved four male participants, two blacks and two whites, applying for entry-level job openings. The auditors formed two teams so that the members of each team were of the same race (the only difference in the application was that one of the testers in each team was assigned a criminal record, a felony drug conviction, and 18 months of prison). The teams applied for 15 jobs per week and the final data included 150 applications by the white pair and 200 by the black pair. The auditors applied for the jobs and advanced as far as they could during the first visit. The application was considered a success only if the auditors were called back for a second interview or hired.

The results showed that 34 per cent of whites with no criminal record were called back, compared with only 17 per cent of those with a criminal record; 14 per cent of blacks without a criminal record were called back, compared to only 5 per cent with a criminal record. Notably, the black auditor without a criminal record received a smaller percentage of callbacks than the white auditor with a criminal record, suggesting the presence of substantial discrimination against blacks in general. Note that Pager found a greater disparity than that found in other audit pair studies in the employment realm. Pager's approach has one notable advantage: the black pair and the white pair were able to use *identical* sets of résumés, which would not have been possible had they been visiting the same employers (the résumés of test partners were similar but not identical). Some have also raised concerns about whether experimenters might have been influenced by the goals of the study to 'find discrimination'. (This is the 'experimenter' effect that Heckman and Siegelman 1993, discuss in the context of the Urban Institute audit studies and that social psychologists have long recognized.)

Bertrand and Mullainathan (2004) also try to measure the extent of race-based labour market discrimination using a slightly different audit strategy that avoids some of the potential pitfalls of direct applicant auditing. Employing a so-called correspondence test methodology, they submitted about 5,000 fictitious résumés in response to employment advertisements appearing in Boston and Chicago newspapers. Their experiment was designed to estimate the racial gap in response rates, measured by phone calls or e-mails requesting an interview. Random application of traditional black or white names to résumés ensures (*a*) that race remains the only component that varies for a given résumé and (*b*) that heterogeneous responses to behaviour or appearance do not affect outcomes (as often occurs with human auditors).

The Bertrand and Mullainathan paper differs from Pager's audit study in that no personal contact with the potential employer takes place in their experiment, so perceived problems with auditor behaviour are eliminated. Bertrand and Mullainathan find significant differences in callback rates for whites and blacks: 'applicants with White names need to send about 10 résumés to get one callback whereas applicants with African–American names need to send about 15 résumés' (Bertrand and Mullainathan 2004, p. 3). Put differently, the advantage of having a distinctly white name translates into roughly eight additional years of experience in the eyes of a potential employer. Whites also appear to benefit much more than blacks from possessing the skills and attributes of a high-quality applicant and from living in a wealthier or whiter neighbourhood. (The difference in callback rates between high and low quality whites is 2.3 percentage points, while for blacks the difference is a meagre one half a percentage point.)

Although these results represent compelling evidence of unlawful discriminatory conduct by employers, the question remains whether the markets are robust enough to reduce or eliminate the apparent disadvantage in the initial hiring process. Fryer and Levitt (2004) indicate that distinctive names do not disadvantage blacks for a variety of adult outcomes. They offer some potential arguments for reconciling their findings with those of Bertrand and Mullainathan (2004). First, if names are considered a noisy initial indicator of race, then they should have no effect once a candidate arrives for the interview. Second, if distinctively black names damage labour market prospects, one might observe more name changes than appear to occur. Finally, with only about ten per cent of jobs being secured through formal résumé-submission processes, the disadvantage of being screened out by certain employers may not be high when other employers and other job search paths remain open.

The combination of the audit studies and the better regression studies seems to tell us that (*a*) there are enough discriminators around for blacks to have to search harder to find employment, (*b*) there are enough non-discriminators around for the resulting unexplained earnings shortfall to be not very high, and (*c*) the unexplained earnings shortfall will overstate discrimination if other legitimate factors are omitted, but will understate the cost of discrimination to blacks because they bear the added search costs of the higher level of employer rejection and any attendant psychological burden that it imposes. To eliminate discrimination would narrow the unexplained earnings gap and remove the added search costs, but this would still leave a substantial unadjusted disparity in black and white earnings.

Statistical Discrimination

A number of theoretical articles have explored whether statistical discrimination contributes to the black–white earnings gap (Arrow 1973; Phelps 1972). This seems unlikely. If, say, blacks are on average treated as their productivity would warrant, then as a class there should be no earnings shortfall, apart from the issue of underinvestment that was discussed above with reference to the Neal and Johnson paper. David Autor and David Scarborough (2004) explore the impact on the hiring and productivity of minority workers, using data from a large nationwide retail firm that changed from an informal worker selection process to one based on standardized testing in 1999. Given that minorities and underprivileged groups on average score lower on such standardized tests, one would expect that this change in the firm's hiring scheme would disadvantage minority workers.

The company originally used informal, paper applications to select candidates for entry level positions. Starting in June 1999, the firm began instituting a computer-based application system that included a personality test for selecting compatible and potentially productive candidates. Autor and Scarborough's sample contains information on test scores, worker demographics, termination date and termination reason (if applicable) for hires made between January 1999 and May 2000 in all the firm's outlets; their sample consists of 34,257 observations. The question they address is how the introduction of testing and the ensuing improvement in the firms' applicant selection procedure affected minority hiring and productivity.

Autor and Scarborough show that if employers statistically discriminate before the test is introduced – that is, if they already use demographic characteristics as a signal for expected productivity of the candidate – then adding testing to the model does not hurt minority hiring but still increases the average productivity of both minority and non-minority workers. The empirical evidence supports this last scenario, revealing uniform increased productivity across demographic groups along with no negative effects on minority hiring.

While we must be careful not to extrapolate the Autor and Scarborough results too far from their context of entry level, near-minimum wage jobs, the paper suggests that before testing was implemented the retail firm either selected workers based on (*a*) some non-race proxy that was correlated (imperfectly) with productivity, or (*b*) statistically discriminated on the basis of race (in violation of federal law), which was itself (imperfectly) correlated with productivity.

The evidence from this one firm confirms the intuition of many economists that statistical discrimination should not be unlawful since on average it should not disadvantage minority workers. One should query, though, whether the legal regime is nuanced enough to legitimize statistical discrimination while prohibiting intentional, animus- based discrimination. Judicial and jury determinations of such issues would presumably be subject to high levels of Type I (incorrectly finding discrimination) and Type II (incorrectly failing to find discrimination) errors.

Sex Discrimination in Employment

Many of the issues discussed above with respect to race discrimination are also relevant to other types of discrimination, including sex discrimination. First, there are questions about whether anti-discrimination law has helped the protected worker. Second, there are issues about whether discrimination can be accurately established. Almost all the groups that seek the aid of anti-discrimination law – minorities, women, the disabled, the elderly – have attributes that non-discriminatory employers might be legitimately concerned about. Under such circumstances, it is difficult to prove that under-representation of any of these groups is caused by discrimination rather than some legitimate factor. The original goal of employment discrimination law in the United States was to eliminate any gap between a worker's productivity and pay caused by discrimination. Today, some argue that the goal of mimicking the outcome of perfectly competitive labour markets is insufficient and that employment discrimination law should more aggressively pursue broader goals of social fairness that will enhance the economic status of disadvantaged groups beyond what a perfect market would provide. According to this view, women should be treated differently to ensure that their role in child-bearing does not disadvantage them in the labour market even if it imposes costs on employers.

Claudia Goldin and Cecelia Rouse (2000) offer an interesting illustration of establishing labour market discrimination in the context of auditions and hiring of musicians for the major US orchestras. To test for sex discrimination in the hiring process, they exploit the changes in the audition process introduced by all major US orchestras in the 1970s and 1980s. Of particular interest for their study was the change to 'blind' auditions, which effectively hid the identity and gender of the applicant from the hiring committee for certain rounds of the audition process. Using audition and roster data spanning several decades and employing an individual fixed effect strategy, they found that the likelihood of female hiring and advancement was increased by the introduction of blind auditions.

More specifically, using audition data from the late 1950s to 1995, Goldin and Rouse found that in blind audition rounds women were as much as 50 per cent more likely to advance from preliminary to final rounds. Furthermore, the likelihood of women winning the finals increased by 33 percentage points if the final round was blind. Using official roster data from 1970 to 1996, they found that completely blind auditions – defined as auditions in which all rounds are conducted with a screen hiding the gender of the applicant – increased the likelihood of a women being hired by 25 per cent. Based on the roster data, blind auditions explain 30 per cent of the increase in female hiring and 25 per cent of the increase in overall female representation in the orchestras. There are, however, some caveats with respect to these findings: first, some estimates have relatively large standard errors that render them statistically insignificant; second, in one scenario – auditions with blind semifinals – the effect on females is persistently strongly negative.

The issue of gender differences in aptitude, specifically aptitude in competitive environments, is explored in an article by Gneezy et al. (2003). Unlike previous studies that tried to explain the gender gap either through occupational self-selection due to differences in abilities and preference or through employer discrimination, Gneezy, Niederle and Rustichini explore the possibility of gender-differentiated performance in competition, which could 'reduce the chance of success for women when they compete for new jobs, promotions, etc'. In a series of controlled

experiments the authors examine the performance of men and women in a computerized maze game as they vary the incentive schemes and group composition for different treatments. They find that, while men receive a significant performance boost in competitive environments such as tournaments, the response of women in competitive environments is more nuanced: they do not significantly change their performance in mixed-sex tournaments, but they do increase their performance in single-sex competitions.

The authors find that under a piece-rate payment scheme men perform only slightly (and not significantly) better than women on average in terms of number of mazes solved. However, when the authors introduce their main competitive treatment of mixed-sex tournaments, they find that men increase their performance significantly, while women's performance remains relatively unchanged.

While women do not seem to receive a performance boost in mixed competitive environments, Gneezy, Niederle and Rustichini also use single-sex tournaments to show that there *are* competitive situations where women increase their performance in response to competition. Both women and men significantly increase their performance in single-sex tournaments, suggesting that women do not dislike competition in general; rather, they dislike competing against men. To explain this, the authors also test for varying feelings of competence across gender. Indeed, once they allow men and women to choose the level of difficulty of the mazes that they are to solve, men choose a higher level of difficulty on average than women do. Whether such factors could explain different pay levels between male and female workers operating under merit pay systems – such as, the lower pay of female stockbrokers, which has been a subject of sex discrimination litigation – is a question that will probably be further explored in the courtroom as well as in the academy.

Conclusion

Anti-discrimination law has generated a number of important social benefits. The elimination of the oppressive race code of the South has been a major benefit of law and policy, opening up all jobs to the most highly qualified candidates. The development of a strong anti-discrimination norm has been an important social asset, and one that merits preservation. To the extent that employers find it natural to be fair to all applicants and workers, the burdens on workers, courts, and employers will be lessened, to the benefit of all.

At the same time, anti-discrimination law has generated some unfortunate unintended consequences, some of which may even threaten the important antidiscrimination norm by undermining its widespread acceptance. I have already alluded to the perverse effects of the situation where an employer might avoid hiring a particular protected worker because of the presence of a governing antidiscrimination law, as some have argued with respect to the protections mandated by the Americans with Disabilities Act. In a regime where the difficulties in ascertaining the existence of discrimination lead to Type I error, firms might find that they are being compelled to hire and compensate certain workers at wages above their levels of productivity. Similarly, as with any negligence-type standard where being adjudicated to have been below a certain level of care can lead to substantial damage awards (including punitive damages), firms have an incentive to take costly measures to be above the threshold that might lead to a finding of discrimination. Tests that may be useful in selecting a high-quality workforce may be avoided if they have, or are thought to have, a disparate impact on certain protected workers that could provide the basis for costly litigation. Note that all these employer adjustments involve costs, but they would appear to involve the benefit of enhancing the employment of groups that are relatively disadvantaged. One might argue that this is a positive development in terms of distributive justice even if it is not actually furthering a corrective justice rationale of eliminating discrimination.

But of course if costs are being imposed on businesses, they will have an incentive to avoid them in the cheapest way possible, which might be through compliance with the legal mandates but could also involve efforts to circumvent the

legal mandates. Indeed, because movements in either direction from the 'non-discriminatory equilibrium' can lead to litigation by whites or blacks or males or females, firms may at times take measures to avoid the litigation risks by using temporary help or by sending their jobs offshore. If these issues were to arise in a racial discrimination context, firms might decide to move offices out to suburban areas or locate where the requirements for hiring black workers would be lessened by the smaller minority benchmark percentages in the relevant labour markets.

As Donohue and Siegelman (1991) noted, the nature of employment discrimination litigation has changed very dramatically in a way that was not anticipated and which may not be entirely desirable. Specifically, most early cases of discrimination complained of failure to hire. These suits tended to open up whole industries or occupations to formerly excluded workers, thereby furthering the objectives of the law. Over time, however, there has been a massive shift in the direction of discharge lawsuits where protected workers claim that they were discriminated against when they were fired. This change sometimes means that low productivity workers can threaten Title VII litigation to hold up an employer for a higher severance package when they are fired for cause. Even worse, firms may find that, at the margin, it is safer not to hire additional protected workers because, at the margin, firms face greater risks from possible, future wrongful discharge discrimination lawsuits than from failure to hire cases. An overall assessment of the impact of anti-discrimination law needs to examine not only the obvious benefits in the form of better treatment of workers through greater professionalization in hiring and human resource management and the productivity enhancements from selecting workers in non-discriminatory ways, but also the array of costs in terms of non-optimal employee selection and retention and firm location decisions, more costly selection processes, and greater litigation costs and legal consulting fees. When every discharge carries the potential for an award of punitive damages, the costs of getting rid of even quite poor workers becomes high. Thus, it may not be surprising that, once the extreme forms of discriminatory conduct were eliminated in the wake of the initial passage of the 1964 Act, further efforts at ratcheting up enforcement of antidiscrimination law seem not to have generated added benefits. Similar arguments about the costs of unintended consequences apply to anti-discrimination enforcement in the realms of mortgage lending, consumer purchases, policing and fighting terrorists.

See Also

- ▶ Black–White Labour Market Inequality in the United States
- ▶ Gender Roles and Division of Labour
- ▶ Jim Crow South
- ▶ Real Wage Rates (Historical Trends)
- ▶ Search Models of Unemployment
- ▶ Social Networks in Labour Markets

Bibliography

Acemoglu, D., and J.D. Angrist. 2001. Consequences of employment? The case of the Americans with Disabilities Act. *Journal of Political Economy* 109: 915–957.

Arrow, K.J. 1973. The theory of discrimination. In *Labor economics*, vol. 4, ed. O.C. Ashenfelter and K.F. Hallock. Aldershot: Edward Elgar.

Autor, D., and D. Scarborough. 2004. *Will job testing harm minority workers?* Working paper, vol. 10763. Cambridge, MA: NBER.

Becker, G. 1957. *The economics of discrimination*. Chicago: University of Chicago Press.

Bertrand, M., and S. Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94: 991–1011.

Burkhauser, R.V., A.J. Houtenville, and L. Rovba. 2006. Accounting for the declining fortunes of working-age people with disabilities. Paper presented at the New York University School of Law Center for Labor and Employment Law Symposium on Assessing the Employment Provisions of the Americans with Disabilities Act, 7 Apr.

Carneiro, P., J. Heckman, and D. Masterov. 2005. Labor market discrimination and racial differences in premarket factors. *Journal of Law and Economics* 48: 1–40.

Chay, K. 1998. The impact of Federal civil rights policy on black economic progress: Evidence from the Equal Employment Opportunity Act of 1972. *Industrial and Labor Relations Review* 51: 608–632.

Conroy, M. 1994. *Faded dreams: The politics and economics of race in America*. Cambridge: Cambridge University Press.
DeLiere, T. 2000. The wage and employment effects of the Americans with Disabilities Act. *Journal of Human Resources* 35: 693–715.
Donohue, J.J. 1994. Employment discrimination law in perspective: Three concepts of equality. *Michigan Law Review* 92: 2583–2612.
Donohue, J.J. 1986. Is Title VII efficient? *University of Pennsylvania Law Review* 134: 1411–1431.
Donohue, J.J. 1989. Prohibiting sex discrimination in the workplace: An economic perspective. *University of Chicago Law Review* 56: 1337–1368.
Donohue, J.J., and J. Heckman. 1991. Continuous versus episodic change: The impact of civil rights policy on the economic status of blacks. *Journal of Economic Literature* 29: 1603–1643.
Donohue, J.J., and P. Siegelman. 1991. The changing nature of employment. *Stanford Law Review* 43: 983–1033.
Freeman, R., R.A. Gordon, D. Bell, and R.E. Hall. 1973. Changes in the labor market for Black Americans, 1948–72. *Brookings Papers on Economic Activity* 1973(1): 67–131.
Fryer, R.G., and S.D. Levitt. 2004. The causes and consequences of distinctively black names. *Quarterly Journal of Economics* 119: 767–805.
Gneezy, U., M. Niederle, and A. Rustichini. 2003. Performance in competitive environments: Gender differences. *Quarterly Journal of Economics* 118: 1049–1074.
Goldin, C., and C. Rouse. 2000. Orchestrating impartiality: The impact of 'blind' auditions on female musicians. *American Economic Review* 90: 715–741.
Heckman, J.J. 1998. Detecting discrimination. *Journal of Economic Perspectives* 12(2): 101–116.
Heckman, J., and B. Payner. 1989. Determining the impact of federal antidiscrimination policy on the economic status of blacks: A study of South Carolina. *American Economic Review* 79: 138–177.
Heckman, J., and P. Siegelman. 1993. The urban institute audit studies: Their methods: Response to comments by John Yinger. In *Clear and convincing evidence: Measurement of discrimination in America*, ed. M. Fix and R. Struyck. Washington, DC: Urban Institute Press.
Jolls, C., and J.J. Prescott. 2004. *Disaggregating employment protection: The case of disability discrimination*, Working paper, vol. 10740. Cambridge, MA: NBER.
Neal, D.A., and W.R. Johnson. 1996. The role of premarket factors in black–white wage differences. *Journal of Political Economy* 104: 869–895.
Orfield, G., and C. Ashkinaze. 1991. *The closing door: Conservative policy and black opportunity*. Chicago: University of Chicago Press.
Oyer, P., and S. Schaefer. 2000. Layoffs and litigation. *RAND Journal of Economics* 31: 345–358.
Oyer, P., and S. Schaefer. 2002a. Litigation costs and returns to experience. *American Economic Review* 92: 683–705.
Oyer, P., and S. Schaefer. 2002b. Sorting, quotas, and the Civil Rights Act of 1991: Who hires when it's hard to fire? *Journal of Law and Economics* 45: 41–68.
Pager, D. 2003. The mark of a criminal record. *American Journal of Sociology* 108: 937–975.
Phelps, E.S. 1972. The statistical theory of racism and sexism. *American Economic Review* 62: 659–661.
Smith, J.P., and F.R. Welch. 1989. Black economic progress after Myrdal. *Journal of Economic Literature* 27: 519–564.

Anti-dumping

Bruce A. Blonigen and Thomas J. Prusa

Abstract

Antidumping is a legal statute that allows for a remedy to offset the effects of dumped imports. Antidumping has emerged as the preferred method of trade protection, accounting for more disputes than all the other trade statutes combined. The economic rationale for current antidumping statutes is weak and generally inconsistent with competition policies. Empirical evidence suggests that antidumping activity is motivated by the same political economy considerations that lead to other forms of trade protection. The economic impact of antidumping remedies can be significant, often dramatically reducing import flows and imposing welfare costs as great as any current trade distortion.

Keywords

Antidumping; Cartels; Collusion; Competition policy; Fair value; Predatory pricing; Price discrimination; Protection; Strategic behaviour; Trade diversion; Voluntary export restraints

JEL Classifications

F13

Antidumping refers to a legal statute that allows for a remedy (typically an import duty) to offset the effects of dumped imports. Under the General Agreement on Trade and Tariffs/World Trade

Organization (GATT/WTO) rules, two tests must be satisfied before a country may impose an antidumping duty on subject imports. First, the imports must be shown to be sold at price that is 'less than fair value'. Second, the dumped imports must be shown to have caused or threaten to cause 'material' injury to a domestic industry.

History and Institutions

The first antidumping (AD) statutes were established in Canada and the United States in the early 1900s. Ultimately, these statutes have been codified into the GATT/WTO statutes. Until the mid-1980s almost all AD activity was confined to four major countries/regions – the United States, the European Union, Australia and Canada (Finger 1993). By the early 1990s countries with newly adopted antidumping statutes accounted for almost one-quarter of AD cases and, since the mid-1990s, new antidumping countries have accounted for well over half of AD complaints (Miranda et al. 1998; Prusa 2001). These new antidumping countries are also far more likely to make an affirmative determination and, consequently, now account for far more than half of all measures in place. Since 1980 GATT/WTO members have filed more complaints under the AD statute than under all other trade laws combined. Worldwide, more AD duties are now levied in any one year than were levied in the entire period from 1947 to 1970.

An antidumping investigation generally proceeds as follow, though there are differences across countries. First, an investigation is initiated when an interested party (often a domestic industry that competes with the imported product) files a petition with the appropriate government agency contending dumping of a particular product(s) from certain import-source countries. The administering government agency (or agencies) then collects data from petitioners and foreign firms that are alleged to be the source of dumped imports and calculates the extent to which imports have been dumped and have injured the domestic industry. Findings of dumping and material injury lead to the imposition of an antidumping duty, which is often equal to the per cent difference between the price of the dumped imports and fair value (that is, the dumping margin). Under WTO statutes, antidumping cases must be reviewed at least every five years to determine whether an antidumping remedy is still appropriate given recent import activity in the subject product.

It is important to understand that antidumping arises from legal concepts. Thus, the meaning of 'less-than-fair-value', causation, and material injury are examined from a legal perspective where previous rulings establish precedence in interpreting the legal definitions. Legal bodies have been active in adjusting these statutes over time. The GATT/WTO antidumping code has undergone significant revisions in nearly every negotiating round, and most countries with these statutes also make periodic legislative changes to their antidumping codes. Many economists have noted that the increase in antidumping activity after these legislative changes is not coincidental. For example, the Tokyo GATT Round contained numerous amendments to the antidumping statute. Of particular importance was the broadening of the definition of the 'less-than-fair-value' concept to capture not only price discrimination, but also sales below cost. Cost-based allegations now account for between one-half and two-thirds of US AD cases (Clarida 1996); an even greater share of EU cases is prosecuted using cost-based methodology (Messerlin 1989).

Given its legal foundation, perhaps it is not surprising that the economic rationale for antidumping statutes is far from clear. A possible rationale is to address predatory pricing practices, where foreign firms are pricing low to induce exit by the domestic firms, allowing monopoly prices in future periods. Economists generally agree that predatory pricing will lead to a welfare loss for a country, but they are sceptical about how often such a strategy is feasible or successful. More importantly, antidumping statutes and practices do not apply the stringent standard used by antitrust (or competition) agencies to determine if pricing is predatory, that is, pricing below marginal cost. Instead, depending on the typical definitions of fair value used by agencies, simple price discrimination across markets or pricing below a level that would return a significant profit to the foreign firm will lead to findings of dumped

imports. Such practices are not generally seen as anticompetitive and, in fact, there is often clear tension between antidumping and competition policy. For example, Staiger and Wolak (1992) have shown that domestic firms can use AD actions to punish foreign firms for refusing to join in collusive actions to raise prices, including the enforcement of price-fixing cartels; examples of price-fixing behaviour in conjunction with AD activity include Ferrovandium and DRAMs. Thus, economists generally believe there is little connection between national welfare considerations and antidumping protection (Stiglitz 1997).

Instead, most economists find evidence that antidumping activity is motivated by the same political economy considerations that lead to other forms of trade protection. While the studies documenting this vary in what proxies they construct to measure political pressure, all find that such non-statutory factors are significant in ultimate antidumping decisions. These studies include Moore (1992); DeVault (1993) and Hansen and Prusa (1996, 1997). Industries with production facilities in politically important districts fare better. There is also some evidence that financial contributions to politicians by industries seeking antidumping protection improve the chance of an affirmative determination. In a related vein, these studies find that antidumping duties are more likely to be levied against particular trading partners. Blonigen and Bown (2003) argue that this finding does not so much reflect a bias against certain countries, but rather reflects that the inability of certain countries to effectively use the threat of retaliation to deter others from using antidumping against it.

In addition, studies of US antidumping activity have found that changes in legal statutes and agency discretion have led to ever greater dumping margins and the likelihood of determining material injury. For example, Hansen and Prusa (1996) show that the US legal change to allow government agencies to consider the all import sources named in an investigation cumulatively (not individually) makes a material injury decision much more likely. This US legal change was later adopted by WTO antidumping statutes in the Uruguay Round and led to both a dramatic increase in the incidence of multi-country cases and also a sharp increase in affirmative determinations (Hansen and Prusa 1996; Tharakan et al. 1998; Irwin 2005). Another example is the documentation by various studies of how the antidumping statutes allow substantial latitude to agencies in how they practically determine dumping margins. Blonigen's (2006) statistical analysis finds that changes in agency discretionary practices is the primary factor behind the rise in average US dumping margins from around 15 per cent in the early 1980s to 60 per cent by 2000.

Direct Economic Effects of Antidumping Statutes and Remedies

The direct economic result of antidumping remedies is to reduce import flows. Such import declines can happen once an investigation is begun and when antidumping remedies are uncertain. In addition, Staiger and Wolak (1994) emphasize that about half the trade impact occurs before the final determination. They argue that trade impact is sufficiently large for the benefits accruing during the investigation to often exceed the costs of filing the petition. Ethier and Fischer (1987); Fischer (1992), Reitzes (1993), and Prusa (1994) also emphasize the dampening impact on trade created by the threat of AD investigation.

From a welfare perspective, a number of studies have documented that domestic firms can gain from such trade-dampening effects, including Hartigan et al. (1989), Blonigen et al. (2004), and Konings and Vandenbussche (2005). However, the latter paper shows that such positive gains are eliminated when foreign firms locate production of the investigated product in country and, thus, avoid the antidumping duties. Prusa (1997) also documents the substantial trade diversion effects that can take place from investigated import sources to non-investigated sources, which provides another reason why such antidumping remedies may not benefit the domestic industry.

Other studies have used computable equilibrium analysis to examine the total welfare consequences of antidumping remedies for a country. As is typical of trade policy welfare analysis, such

losses to consumers are typically estimated to outweigh the gains to the protected producers for antidumping protection. For example, using a computable general equilibrium model, Gallaway et al. (1999) estimate that the cumulative effect of all antidumping duties in place leads to an annual four billion dollar welfare loss for the United States. This figure places this form of trade protection as second only to the restrictive and comprehensive quotas on textiles and apparel (Multifiber Arrangement) in terms of welfare costs.

Indirect Economic Effects of Antidumping Statutes and Remedies

Beyond these typical trade and welfare considerations, economists have pointed to a number of features of antidumping programmes that may cause a greater range of ancillary (or indirect) effects that are often unique to this form of trade protection. In fact, this is where the bulk of recent economic literature has centred its attention, and insights often come from thinking about strategic considerations applying game theoretic techniques.

Such issues are pervasive in analysing the decision to file an antidumping case and its likely chance of success. A foreign industry can almost guarantee it will not be subject to antidumping duties if it charges sufficiently high prices in its export markets. On the other hand, a domestic industry has incentives to look 'weak' to make an injury determination more likely, which could lead it to charge higher prices (produce less) than optimal, or lay off more workers than it otherwise would. Ethier and Fischer (1987); Fischer (1992), Reitzes (1993), and Prusa (1994) are examples of applied game theory pieces that document these possible strategic decisions by domestic and foreign firms to influence future antidumping outcomes. Anderson (1992, 1993) examines the potential interdependence of antidumping with another form of trade protection: voluntary export restraints (VERs). The artificial scarcity created by the VERs generates rents for foreign firms that are typically divided up by their market shares. This perversely gives the foreign firms incentives to 'dump' their products to garner larger market shares, which makes antidumping investigations and remedies more likely.

The strategic interactions described above are non-cooperative in nature, but a number of papers have examined how antidumping can elicit various forms of cooperative strategic behaviour. These studies primarily provide theoretical analysis, showing how antidumping law can facilitate or sustain collusive cartel pricing by foreign and domestic firms; such studies include Staiger and Wolak (1989); Prusa (1992), and Veugelers and Vandenbussche (1999). Taylor (2004) and Zanardi (2004) provide empirical examinations of collusive behaviour in antidumping activity using US data.

Strategic interactions surrounding antidumping petitions may also occur amongst domestic firms. Cassing and To (2004) show that the decision by a domestic firm to join an antidumping petition can signal its efficiency to other firms in the market. Thus, for example, some domestic firms may not join a petition to signal to others that they have low costs.

Once antidumping remedies are in place, other strategic reactions are possible too. As mentioned above, a foreign firm can 'jump' the antidumping duties and relocate its production to either the domestic market or to a third country that is not subject to the duties. Belderbos (1997) and Blonigen (2002) document significant tariff-jumping of antidumping duties in Europe and the United States. Interestingly, if foreign firms differ in their ability to make such investments, then antidumping might particularly burden firms who cannot make such adjustments. Ironically, this means the foreign firms who are most able to 'jump' the AD duty potentially have an incentive to encourage antidumping actions (Blonigen and Ohno 1998).

The ability of firms to reduce their antidumping duties in subsequent administrative reviews also provides interesting incentives to firms. Such reviews examine recent data to recalculate antidumping duties, which creates a dynamic environment for price setting by the foreign firm. Blonigen and Park (2004) develop a model of dynamic pricing decisions by foreign firms facing the possibility of antidumping duties and subsequent recalculations in future periods. They first

show that, if antidumping duties are a certainty when a foreign firm dumps, then the only firms that will dump care very little about the future (high discount rates). Over time the punitive antidumping duties will cause them to dump even more. However, if antidumping remedies are uncertain, foreign firms that have *ex ante* low expectations of antidumping remedies will quickly reduce their dumping once, to their surprise, they become subject to antidumping duties. Blonigen and Park confirm these hypotheses using data on US antidumping investigations. In a related paper, Blonigen and Haynes (2002) find that foreign firms subject to antidumping duties alter their behaviour to fully pass through exchange rate changes and also pass through greater than 100 per cent of the antidumping duty onto the prices in their export market.

Blonigen and Prusa (2003) provide a detailed review of the economics literature on antidumping and also point towards what they consider fruitful areas for future research. These include the treatment of antidumping in competition policy, effects on downstream industries and import/export companies, and comparisons of antidumping statutes across various WTO member countries. The U.S. Antidumping and Countervailing Duty Database and the Global Antidumping Database should play an important role in facilitating future research in antidumping.

See Also

- ▶ International Trade Theory
- ▶ Tariffs
- ▶ Trade Costs

Bibliography

Anderson, J.E. 1992. Domino dumping I: Competitive exporters. *American Economic Review* 82: 65–83.

Anderson, J.E. 1993. Domino dumping II: Anti-dumping. *Journal of International Economics* 35: 133–150.

Belderbos, R.A. 1997. Antidumping and tariff jumping: Japanese firms. DFI in the European Union and the United States. *Weltwirtschaftliches Archiv* 133: 419–457.

Blonigen, B.A. 2002. Tariff-jumping antidumping duties. *Journal of International Economics* 57: 31–50.

Blonigen, B.A. 2006. Evolving discretionary practices of U.S. antidumping activity. *Canadian Journal of Economics* 39: 874–900.

Blonigen, B.A., and Y. Ohno. 1998. Endogenous protection, foreign direct investment, and protection-building trade. *Journal of International Economics* 46: 205–227.

Blonigen, B.A., and S.E. Haynes. 2002. Antidumping investigations and the passthrough of exchange rates and antidumping duties. *American Economic Review* 92: 1044–1061.

Blonigen, B.A., and C.P. Bown. 2003. Antidumping and retaliation threats. *Journal of International Economics* 60: 249–273.

Blonigen, B.A., and T.J. Prusa. 2003. Antidumping. In *Handbook of international economics*, ed. E. Kwan Choi and J. Harrigan. Malden: Blackwell.

Blonigen, B.A., and J.-H. Park. 2004. Dynamic pricing in the presence of antidumping policy: Theory and evidence. *American Economic Review* 94: 134–154.

Blonigen, B.A., K. Tomlin, and W.W. Wilson. 2004. Tariff-jumping FDI and domestic firms' profits. *Canadian Journal of Economics* 37: 656–677.

Cassing, J., and T. To. 2008. Antidumping, signaling and cheap talk. *Journal of International Economics* 75: 373–382.

Clarida, R.H. 1996. Dumping in theory, in policy, and in practice. In *Fair trade and harmonization*, ed. J. Bhagwati and R. Hudec. Cambridge, MA: MIT Press.

DeVault, J.M. 1993. Economics and the international trade commission. *Southern Economic Journal* 60: 463–478.

Ethier, W.J., and R.D. Fischer. 1987. The new protectionism. *Journal of International Economic Integration* 2: 1–11.

Finger, J.M. 1993. *Antidumping: How it works and who gets hurt*. Ann Arbor: University of Michigan Press.

Fischer, R.D. 1992. Endogenous probability of protection and firm behavior. *Journal of International Economics* 32: 149–163.

Gallaway, M.P., B.A. Blonigen, and J.E. Flynn. 1999. Welfare costs of US antidumping and countervailing duty laws. *Journal of International Economics* 49: 211–244.

Global Antidumping Database. Online. Available at: http://econ.worldbank.org/ttbd/gad/. Accessed 12 May 2007.

Hansen, W.L., and T.J. Prusa. 1996. Cumulation and ITC decision making: the sum of the parts is greater than the whole. *Economic Inquiry* 34: 746–769.

Hansen, W.L., and T.J. Prusa. 1997. The economics and politics of trade policy: An empirical analysis of ITC decision making. *Review of International Economics* 5: 230–245.

Hartigan, J.C., S. Kamma, and P.R. Perry. 1989. The injury determination category and the value of relief from dumping. *The Review of Economics and Statistics* 71: 183–186.

Irwin, D.A. 2005. The rise of U.S. antidumping activity in historical perspective. *The World Economy* 28: 651–668.
Konings, J., and H. Vandenbussche. 2005. Antidumping protection and markups of domestic firms. *Journal of International Economics* 65: 151–165.
Messerlin, P.A. 1989. The EC antidumping regulations: A first economic appraisal, 1980–85. *Weltwirtschaftliches Archiv* 125: 563–587.
Miranda, J., R.A. Torres, and M. Ruiz. 1998. The international use of antidumping: 1987–1997. *Journal of World Trade* 32: 5–71.
Moore, M.O. 1992. Rules or politics? An empirical analysis of ITC anti-dumping decisions. *Economic Inquiry* 30: 449–466.
Prusa, T.J. 1992. Why are so many antidumping petitions withdrawn? *Journal of International Economics* 33: 1–20.
Prusa, T.J. 1994. Pricing behavior in the presence of antidumping law. *Journal of Economic Integration* 9: 260–289.
Prusa, T.J. 1997. The trade effects of US antidumping actions. In *The Effects of US Trade Protection and Promotion Policies*, ed. R.C. Feenstra. Chicago: University of Chicago Press.
Prusa, T.J. 2001. On the spread and impact of antidumping. *Canadian Journal of Economics* 34: 591–611.
Reitzes, J.D. 1993. Antidumping policy. *International Economic Review* 34: 745–763.
Staiger, R.W. and Wolak, F.A. 1989. Strategic use of antidumping law to enforce tacit international collusion. Working paper no. 3016. Cambridge, MA: NBER.
Staiger, R.W., and F.A. Wolak. 1992. The effect of domestic antidumping law in the presence of foreign monopoly. *Journal of International Economics* 32: 265–287.
Staiger, R.W. and Wolak, F.A. 1994. Measuring industry specific protection: Antidumping in the United States. *Brookings Papers on Economic Activity Microeconomics* 1994, 51–118.
Stiglitz, J.E. 1997. Dumping on free trade: The US import trade laws. *Southern Economic Journal* 64: 402–424.
Taylor, C.T. 2004. The economic effects of withdrawn antidumping investigations: is there evidence of collusive settlements? *Journal of International Economics* 62: 295–312.
Tharakan, P.K.M., D. Greenaway, and J. Tharakan. 1998. Cumulation and injury determination of the European community in antidumping cases. *Weltwirtschaftliches Archiv* 134: 320–339.
U.S. Antidumping and Countervailing Duty Database. Online. Available at: http://darkwing.uoregon.edu/Bbruceb/adpage.html. Accessed 12 May 2007.
Veugelers, R., and H. Vandenbussche. 1999. European anti-dumping policy and the profitability of national and international collusion. *European Economic Review* 47: 1–28.
Zanardi, M. 2004. Antidumping law as a collusive device. *Canadian Journal of Economics* 37: 95–122.

Anti-poverty Programmes in the United States

Robert A. Moffitt

Abstract

Economic theory suggests that the extent of redistribution should be constrained by its direct and indirect costs, including disincentive effects. The emphasis in the United States has been on programmes that emphasize employment as well as in-kind rather than cash redistribution, and that provide benefits to populations with special needs. Research on their effects has shown them to decrease poverty rates and the poverty gap but to have labour-supply disincentives as well. Reforms to the main cash programme in the 1990s have increased earnings and employment.

Keywords

Aid to Families with Dependent Children (AFDC) (USA); Altruism; Anti-poverty programmes in the United States; Child care subsidies; Crowding out; Earned Income Tax Credit (USA); Food stamps; Free-rider problem; Head Start (USA); In-kind transfers; Job Corps (USA); Labour supply; Low-income housing policy; Marginal utility of consumption; Means-tested transfers; Medicaid (USA); Negative income tax; Poverty gap; Supplemental Security Income (SSI) (USA); Temporary Assistance for Needy Families (TANF) (USA); Working Families Tax Credit (UK)

JEL Classification

H5

Anti-poverty programmes in the United States have received much attention from the economics profession since the 1970s. Economists have studied their effectiveness in reducing poverty and increasing well-being among the poor, their rationale and goals, and trends in their caseloads and

expenditures. Scholars have also extensively studied the effects of anti-poverty programmes on a wide range of individual and family behaviours.

Rationale and Design Issues

Anti-poverty programmes are generally considered to arise from altruism on the part of non-poor voters, who wish to transfer resources, for charitable reasons, to those who have low incomes or assets. Such charitable support is generally considered to be suboptimally provided if left to the private sector because of the free-riding problem that arises when one individual's contribution to the poor makes other givers better off, so individuals have an incentive to let others contribute rather than contribute themselves.

However, the exact nature of the preferences of the non-poor – let us call them voters, since the United States is a democracy – are not well understood. In the classic utilitarian model, the social welfare function equals the sum of individual utilities and the marginal utility of income is assumed to decline with income, so that a dollar redistributed from a high-income person to a low-income person raises social utility. One issue with this framework is whether the 'weights' that the voters assign to the poor are the same as marginal utility of income weights, and today most analysts assume those weights to deviate in an arbitrary way and to simply reflect voter preferences that will vary from group to group and from country to country. Another important distinction is whether the voters desire to increase the utility of the poor *per se*, as the utilitarian model implies, or to increase their consumption of specific goods like food, housing, and medical care. Redistributing in the latter fashion, resulting in what are termed 'in-kind' transfers, is quite common in practice, and economists have often assumed that it implies that voters are paternalistic in the sense that they wish to override the spending preferences of the poor themselves. Redistributing purely in the form of income, for example, would allow recipients to allocate the transfer in a way that maximizes their utility as they see it. Another rationale for in-kind transfers is that they induce only those with the highest marginal utility of consumption of those goods to accept such transfers, which induces a desirable (from the voter's point of view) selection from the low-income population to those who need it most (Nichols and Zeckhauser 1982; Blackorby and Donaldson 1988), and yet another is that they reduce the incentive of the recipient to alter behaviour to increase later transfers (Bruce and Waldman 1991).

Whatever the preferences of the voters, the main issue in models of optimal provision of anti-poverty benefits to the poor is the trade-off between the benefits of redistribution and the direct and indirect costs of the transfer. The direct costs arise because taxation has its own resource cost and the indirect costs arise because the transfer distorts the behaviour of the recipients. As in the classic models of taxation, lump-sum transfers are not possible and so transfers alter the prices of various goods in the utility function. In the well-known Mirrlees (1971) model, the main margin examined is work effort, which is reduced by transfers, and optimal redistribution proceeds up to the point where the marginal benefits of additional redistribution are counterbalanced by the marginal losses arising from reductions in work effort.

However, one of the main areas of research on anti-poverty programmes, particularly those that are empirical in nature, has been on other possible margins of adjustment by programme recipients. Transfers may reduce incentives to invest in human capital, reduce incentives to save if assets are taxed by the programme, increase incentives to have additional children if benefits are tied to family size, change incentives to marry if marital status affects benefits, or increase incentives to migrate from one jurisdiction to another to obtain higher benefits if benefits vary within a country. For in-kind programmes, there is also potential 'leakage' in the consumption effects. For example, giving a family either a lump-sum amount of food or a subsidy to the price of food may lead them to reduce their own expenditures on food in order to spend more on other consumption items.

The prototype of a transfer programme that aims to balance redistribution and disincentives

is the negative income tax (NIT) (Watts 1987). In an NIT, recipients who have no income receive a maximal benefit but the size of the transfer declines as income rises. Thus those with lower incomes receive greater benefits than those with higher incomes, as most models imply should occur, but the rate at which benefits are reduced as income rises is generally taken to be less than 100%.

This provides some incentive to work, and work disincentives are therefore controlled by the rate of benefit reduction. A transfer system with a 100% reduction rate, particularly one that extends relatively high into the income distribution, is said to create a 'poverty trap' because individuals cannot escape poverty through modest increases in income. The first formal demonstration of the optimality of an NIT was provided again by Mirrlees (1971), who showed that such a programme results from an optimal utilitarian model. This general paradigm applies to the other margins mentioned above as well, for in each case a programme can be designed to provide the highest benefits to those with the lowest resources while paying attention to the effect of the programme on the price of changing behaviour (undertaking human capital investment, saving, and so on). An important modification of the Mirrlees models appears in Diamond (1980) and Saez (2002), who showed that consideration of the 'extensive' margin of work – namely, the decision to work at all rather than the decision of how many hours to work, which was the focus in the Mirrlees model – can lead to earnings subsidies, where the marginal 'tax rate' on earnings at the bottom of the income distribution is negative rather than positive for some range. The Earned Income Tax Credit in the United States and the Working Families Tax Credit in the United Kingdom are important examples of such earnings subsidies.

Finally, a benefit-provision issue that economists have studied is the relative merits of redistribution by a central government versus local governments within a country. For many years it was assumed that the utility of the poor in all jurisdictions should affect the utility of voters in all jurisdictions equally, which leads to a central government programme. But Pauly (1973) and others have argued that local voters care more about the poor in their own jurisdictions, making redistribution partly a local public good, although they may care to some extent about the poor in other jurisdictions as well. This leads to a mixed central–local system in which the central government subsidizes local governments because of the limited interest of all voters but allows localities to spend on redistribution out of their own resources as well. This leads to subsidy mechanisms such as block grants, matching grant programmes and related funding mechanisms. This structure is found in the United States but also in some European countries.

Anti-poverty Programmes

There are a large number of anti-poverty programmes in the United States whose structure and expenditure have changed over time (Moffitt 2003). We shall ignore Social Security, which has a major impact on poverty rates of the elderly but which is generally considered to be a social insurance programme rather than a means-tested transfer programme. The most well-known and heavily studied programme, and that which historically most resembled an NIT, is the Temporary Assistance for Needy Families (TANF) programme, which was called the Aid to Families with Dependent Children (AFDC) programme prior to 1996. The TANF programme provides monthly cash benefits to families with low income and assets, but primarily to those headed by a single parent (mostly single mothers). The benefit-reduction rate in the programme varies across states but is most often around 50%. However, the TANF programme also has some non-NIT features – specifically, it has work requirements that mandate that most able-bodied parents work at least some minimum number of hours per week as a condition of receiving benefits, and time limits, which stipulate that parents can receive benefits for only a limited number of years over their lifetimes. These latter provisions were enacted in 1996.

While the AFDC programme was one of the leading US anti-poverty programmes in the 1960s and 1970s, when its caseloads and expenditures were among the largest of US programmes, in 2007 it ranked only sixth in terms of expenditure and fifth in terms of caseload (Moffitt 2007). It is smaller than the Medicaid programme, which provides medical subsidies to the poor; the Supplemental Security Income (SSI) programme, which provides benefits to poor families with aged adults and disabled adults and children; the Earned Income Tax Credit (EITC), which provides tax credits to working families; Food Stamps, which provides food subsidies to the poor; and housing programmes for the poor. Per capita expenditures on AFDC–TANF have steadily declined since the late 1970s, whereas those on the other programmes have grown by amounts much greater in magnitude. In 2007, total real per capita expenditures in the largest means-tested transfer programmes in the United States had more than quadrupled since 1968 and had grown by 60% just since 1990 as a result of the growth in many of these programmes.

The Medicaid programme, the largest programme in the United States, is a diverse programme covering several different populations. The four primary groups served are low-income single mothers and their children; the low-income elderly; the low-income disabled; and individuals in nursing homes or long-term care with low income and assets. Expenditures and caseloads in the programme grew rapidly in the late 1980s and early 1990s as a result of expansions of eligibility for low-income mothers and children and growth of disabled recipients, and have continued to grow secularly because of growth in the demands for long-term care of the elderly. The United States does not have national health insurance and the size and growth of the Medicaid programme partially reflects that fact. With a few exceptions in certain parts of the programme, there is no benefit-reduction rate in the programme; either the full package of benefits is provided or none at all.

The SSI programme pays cash benefits to low-income individuals who are blind or disabled, and to the low-income elderly. The programme also saw very rapid growth in the early 1990s as a result of increases in disabled, child, and non-citizen recipients. The definition of disability for adults is quite stringent; 60% of applications are denied. The disability definition for children is more elastic and has fluctuated in stringency over time. The programme has a nominal 50% benefit-reduction rate.

The EITC also grew rapidly in the late 1980s and early 1990s, while the Food Stamp programme grew most rapidly after its introduction in the late 1960s and early 1970s, but also most recently (since 2000). The EITC has a subsidy rate of up to 40% and a maximum clawback rate of 21%, while the Food Stamp programme has a nominal 30% benefit-reduction rate.

Other important programmes include those covering housing, child care and training programmes. Housing programmes, which have a typical benefit-reduction rate of 30%, grew most rapidly in the late 1970s and early 1980s, and have seen only modest growth since that time. Child care subsidies in the United States are spread over several different programmes serving overlapping populations, including the welfare poor but also the 'working' poor. Expenditures have grown modestly since 2000 as the need for employment support has become increasingly recognized. Included in the child care framework is the Head Start programme, whose goal is to assist child development in pre-school children of low-income families but which also serves a child care function. The United States spends relatively little on training programmes, and has changed the name and nature of its programme for adults several times since the 1970s in an attempt to make the programmes more effective. Perhaps the most popular programme is the Job Corps, a high-cost residential-based programme for disadvantaged young men and women.

Several patterns can be discerned in the US transfer programme system. First, in-kind transfers are preferred to cash transfers. The only programme that is a pure cash transfer programme is the AFDC–TANF programme, which has declined in importance because of its unpopularity and is now coupled with work requirements in any case. The most popular programmes are those

that subsidize medical and food expenditures; those which subsidize housing and child care expenditures are large as well. Second, subsidies that serve specialized populations with specific identifiable needs are preferred to subsidies based on low income *per se*. The SSI programme, which is cash in nature, is the best example of this preference. However, even the EITC could be argued to fit this category, for it provides cash but only to a specific population viewed as meritorious, namely, low-wage workers. Third, an increasing emphasis on employment is apparent. The EITC reflects this emphasis as do the recent reforms in the AFDC–TANF programme and increases in child care subsidies. Fourth, US voters dislike providing subsidies to low-income single-mother families, who are viewed unfavourably because of US views towards marriage. All four of these features are in explicit conflict with the original idea of an NIT as espoused by Milton Friedman, Robert Lampman, James Tobin, and others, who saw the ideal transfer programme as one that provided only cash benefits, on the basis of income only, and without preference for family structure or type.

Research Findings on the Effects of US Anti-poverty Programmes

One overriding issue of interest in research on US anti-poverty programmes is whether such programmes have, in fact, reduced poverty. The evidence indicates that they have (Scholz and Levine 2001). In 1997, the system of means-tested transfer programmes in the United States reduced the poverty rate of families from 29 to 26%, a modest amount. However, the programmes also raised the incomes of many poor families even if not by enough to cross the poverty line, for the programmes filled in 27% of the poverty gap (defined as the total dollar gap between the poverty line and the incomes of poor families). The most important programme in reducing poverty was Medicaid; SSI and the EITC were also important. It is often noted that these estimates should be considered to be an upper bound for the true effect of transfer programmes on poverty because the work disincentives of the programmes themselves cause a reduction in income, which widens the poverty gap and increases the poverty rate to some offsetting extent.

In addition to this issue, there has been a very large amount of research on the behavioural effects of US anti-poverty programmes. By far the most research has been conducted on the AFDC–TANF programme, where the primary focus prior to 1996 was on its effects on labour supply, marriage and fertility, and a few other behaviours (Moffitt 1992). Most research on labour supply indicated, as economic theory would predict, negative effects of the programme as a whole. However, the effects of reducing the benefit-reduction rate have been shown to be mostly zero or negligible, with the general interpretation being that such changes bring in new recipients who experience labour-supply reductions that offset the labour supply increases of those initially on the programme. Research on marriage and fertility effects of AFDC has shown mostly small but non-zero effects in reducing marriage and increasing childbearing. Research conducted on the effects of the 1996 reform of the programme (Blank 2002; Moffitt 2003; Grogger and Karoly 2005) has shown the reform, whose major elements were work requirements and time limits, to have had positive effects on average employment, earnings, and family income and negative effects on welfare usage. However, some research also suggests that there is a group of very disadvantaged families who were made worse off by the reform. The research also has shown the reform to have had little if any effect on marriage and fertility behaviour and to have had modest effects, if any, on children in low-income families.

There has been a fair amount of research on other programmes as well. The Medicaid programme appears to have modest negative effects on labour supply and expansions in the programme have led to ‘crowdout’ of private health insurance, but the programme has also been shown to have had many favourable effects on health, particularly that of children (Gruber 2003). Research on the SSI programme has focused particularly on reasons for fluctuations

in the size of the caseload, but has also concerned work incentives, where both benefit-reduction rates and other employment-incentive programmes have been shown to have had little effect (Daly and Burkhauser 2003). Research on child care programmes have shown them to have had positive effects on female employment, and Head Start has been shown to have some positive effects on child outcomes, but which fade out over time (Blau 2003). Work on training programmes has shown them to have different effectiveness for different groups, with several low-cost programmes found to be effective in increasing earnings for single mothers and with the high-cost Job Corps programme found to be effective for disadvantaged youth, but with no type of programme having been found to have a significantly positive rate of return for adult men (Lalonde 2003).

See Also

- ▶ Nutrition and Public Policy in Advanced Economies
- ▶ Poverty Alleviation Programmes
- ▶ Taxation and Poverty
- ▶ Welfare State

Bibliography

Blackorby, C., and D. Donaldson. 1988. Cash versus kind, self-selection, and efficient transfers. *American Economic Review* 78: 691–700.

Blank, R. 2002. Evaluating welfare reform in the United States. *Journal of Economic Literature* 40: 1105–66.

Blau, D. 2003. Child care subsidy programs. In *Means-tested transfer programs in the United States*, ed. A.R. Moffitt, 443–516. Chicago: University of Chicago Press.

Bruce, N., and M. Waldman. 1991. Transfers in kind: Why they can be efficient and nonpaternalistic. *American Economic Review* 81: 1345–51.

Daly, M., and R. Burkhauser. 2003. The supplemental security income program. In *Means-tested transfer programs in the United States*, ed. A.R. Moffitt. Chicago: University of Chicago Press.

Diamond, P. 1980. Income taxation with fixed hours of work. *Journal of Public Economics* 13: 101–10.

Grogger, J., and L. Karoly. 2005. *Welfare reform: Effects of a decade of change*. Cambridge, MA: Harvard University Press.

Gruber, J.D. 2003. Medicaid. In *Means-tested transfer programs in the United States*, ed. A.R. Moffitt. Chicago: University of Chicago Press.

Lalonde, R. 2003. Employment and training programs. In *Means-tested transfer programs in the United States*, ed. A.R. Moffitt. Chicago: University of Chicago Press.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* 38: 175–208.

Moffitt, R. 1992. Incentive effects of the U.S. welfare system: A review. *Journal of Economic Literature* 30: 1–61.

Moffitt, R. (ed.). 2003. *Means-tested transfer programs in the United States*. Chicago: University of Chicago Press.

Moffitt, R. 2007. Four decades of anti-poverty policy: past developments and future directions. In *Focus*. Institute for Research on Poverty, University of Wisconsin (forthcoming).

Nichols, A., and R. Zeckhauser. 1982. Targeting transfers through restrictions on recipients. *American Economic Review* 78: 240–4.

Pauly, M. 1973. Income redistribution as a local public good. *Journal of Public Economics* 2: 35–58.

Saez, E. 2002. Optimal income transfer programs: Intensive versus extensive labor supply responses. *Quarterly Journal of Economics* 117: 1039–73.

Scholz, J.K., and K. Levine. 2001. The evolution of income support policy in recent decades. In *Understanding poverty*, ed. S. Danziger and R. Haveman. New York: Russell Sage.

Watts, H. 1987. Negative income tax. In *The new palgave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman. London: Macmillan.

Anti-trust Enforcement

Joseph E. Harrington

Abstract

This article explores the enforcement of those laws intended to promote competitive markets through the prohibition of certain practices such as price-fixing, welfare-reducing mergers, and monopolization. The discovery and prosecution of violations are examined, including the role of leniency programmes. The determination of penalties is investigated with an assessment of their relationship to optimal penalties. Enforcement policy is found to vary over time and its

determinants are reviewed. Finally, the efficacy of enforcement is assessed.

Keywords

Antitrust enforcement; Antitrust penalties; Cartels; Collusion; Corporate Leniency Program; Price fixing

JEL Classifications

L40

Antitrust enforcement is the process whereby a more competitive environment is created through the prohibition of certain practices deemed illegal by antitrust laws.

Restraints of trade such as price-fixing and bid-rigging are prohibited in the United States under section 1 of the Sherman Act of 1890 and in the European Union under article 81 of the Treaty of the European Communities of 1999. Practices designed to create monopolies (such as predatory pricing and tying) are prohibited in the United States under section 2 and in the European Union under article 82. Mergers that are harmful to competition are prohibited in the United States under section 7 of the Clayton Act of 1914 and in the European Union under article 2(3) of the Merger Regulation. Although this article adopts a US focus, much of what is described is applicable to many OECD countries. (For a more general treatment on antitrust policy, see Motta 2004, for the European Union and Viscusi et al. 2005, for the United States.)

Detection of Antitrust Offences

Enforcement can involve three stages: (*a*) discovery and evaluation of a possible antitrust violation; (*b*) prosecution when it is deemed there is a violation; and (*c*) levying of penalties and enacting of remedies when prosecution is successful.

Antitrust cases can arise in a variety of ways. With a recent exception noted below, cartels are generally discovered not by the antitrust authorities but rather by customers, employees, and even competitors. Though not yet widely used, economic and econometric methods for detecting collusion include determining whether: (*a*) firm behaviour is inconsistent with competition; (*b*) there is a structural break in behaviour; (*c*) the behaviour of suspected colluding firms differs from that of some benchmark competitive firms; and (*d*) a collusive model fits the data better than a competitive model (Harrington 2006). In contrast, prospective merger cases are brought by the participants themselves to the antitrust authorities, as mandated by the Hart–Scott–Rodino Act of 1976. In evaluating a proposed merger, the primary considerations are the extent to which it would raise price and whether there are offsetting cost savings.

Antitrust Penalties

In the case of price-fixing, the government levies fines at the corporate level which, as a result of the Sentencing Reform Act of 1984, can be as high as twice the gross pecuniary gain of the defendant or twice the pecuniary loss of the victims (though a Supreme Court decision in 2005 has since put these guidelines into jeopardy). The most significant financial penalty comes from private damages which, due to the Clayton Act, allow direct buyers to receive compensation equal to three times the damages. At the individual level, the government imposes fines and prison sentences; since 1970, 53 per cent of convicted individuals have been imprisoned (Gallo et al. 2000). The use of government fines is common in many other countries, although prison sentences and civil damages are unique to the United States and Canada.

Are these penalties optimal? An optimal penalty is one that deters only those activities that are welfare-reducing. If the gain to the offenders is g, the loss to other agents is l, the probability of being penalized is p, and the penalty is f then optimality requires: $g - pf \geq 0$ if and only if $g \geq l$ (Polinsky and Shavell 2000). Therefore, the optimal penalty is $f = l/p$ In practice, private damages are calculated as $(P^c - P^{bf})Q^c$ where P^c is the observed (collusive) price, Q^c is the number

of units sold, and P^{bf} is the 'but for' price, that is, the price that would have been charged but for collusion. $P^c - P^{bf}$ is referred to as the 'overcharge'. A major source of contention in many price-fixing cases is the determination of P^{bf}, for which reducedform estimation methods are largely deployed with the use of data encompassing both the cartel and non-cartel regimes. The 'before and after' approach is quite common and entails estimating: $P(t) = \delta + \beta X(t) + \gamma v(t) + \varepsilon(t)$ where $P(t)$ is price, $X(t)$ is a vector of demand and cost shifters, and $v(t)$ is a dummy variable that equals one in those periods that firms were colluding (Page 1996). If $\widehat{\delta}$ and $\widehat{\beta}$ are the parameter estimates, then $P^{bf}(t) = \widehat{\delta} + \widehat{\beta}X(t)$. Since damages, as calculated in practice, ignore deadweight loss, penalties are neither optimally punitive nor compensatory: $g < (P^c - P^{bf})Q^c < l$. Government fines also suffer from this deficiency as they tend to be proportional to sales, P^cQ^c.

Of course, if collusion serves only to reduce supply, then $l > g$ and thus we should prevent all collusion, in which case $f \geq g/p$ is desired. As cartels continue to form, penalties clearly fall short. But how far away are they from being an effective deterrent? In practice, cases are largely settled out of court and single (not treble) damages are typical (Lande 1993). For international cartels over 1990–2003, Connor (2004) calculates private and public recovery in the United States was only 115 per cent of damages. Bryant and Eckard (1991) infer from observed cartel lengths that the chances of a price-fixing cartel being indicted in a 12-month period is 11–15 per cent. Though that estimate relies on a properly specified functional form for the distribution on cartel lifetimes, it is safe to say that the probability of a cartel being discovered and paying penalties is well below one, so that financial penalties are woefully inadequate. What may be more effective is the use of prison sentences (Werden and Simon 1987).

Although remedies have been used in price-fixing cases (for example, a ten-year consent decree in 1994 placed restrictions on announcements of future price changes by airlines), they are typically more important in merger and monopolization cases. Some proposed mergers receive government approval only after restructuring, such as the selling of assets that, if retained by the newly merged firm, would significantly harm competition. In rare cases, the authorities seek to prevent the merger entirely. In the case of monopolization, remedies may be either behavioural or structural. Behavioural remedies could, for example, require a firm to license intellectual property to competitors (as with Xerox) or prohibit certain contractual arrangements (as with Microsoft). Structural remedies are typically quite draconian and accordingly rare. Notable examples include the break-up of Standard Oil in 1911 and AT&T in 1984. A lower court initially ordered Microsoft to be broken into two companies – one with the operating system and the other with applications – though it was later remanded by the US Court of Appeals, and the Department of Justice (DOJ) stopped pursuing it as a remedy.

Corporate Leniency Program

One of the most significant innovations in antitrust enforcement in recent years is the 1993 revision of the DOJ's Corporate Leniency Program and the institution of a similar programme by the European Commission in 1996. The first member of a cartel to come forward and cooperate receives full amnesty with respect to government penalties and liability for only single damages. As a condition of entering the programme, company representatives must answer an 'omnibus question' which asks them whether they know of any collusion in other markets. Failure to truthfully answer that question results in the loss of all amnesty. This policy has proven useful for both the discovery and the prosecution of cartels.

Under the standard repeated game framework, a leniency programme affects the stability of collusion through the usual equilibrium condition: the expected payoff from continuing to collude must be at least as great as the payoff to a firm from (optimally) cheating on the cartel. (The discussion here is based on Harrington, 2005;

see also Motta and Polo 2003, and Spagnolo 2003.) More leniency enhances the payoff to cheating because a firm that does so can simultaneously apply for amnesty and thereby reduce expected penalties. However, leniency also affects the expected collusive payoff because firms anticipate the possibility of using the programme in the future. More leniency lowers penalties in the event that leniency is received and thus can raise the payoff from continuing to collude. But it is also possible that waiving a higher fraction of penalties *increases* future expected penalties. The reason is that there can be two equilibria: one in which all firms apply for amnesty and one in which none does. The latter can Pareto-dominate because only one firm can receive amnesty and use of the programme results in certain conviction. More leniency can destabilize the Pareto-preferred equilibrium in which all firms refrain from using the programme because it becomes too attractive for a firm to apply (given that other firms do not). Although there are then several countervailing forces, it is generally optimal to provide some leniency, and conditions are not too restrictive for it to be optimal to waive all penalties.

Intensity of Antitrust Enforcement

An enforcement policy is described not just by the types of cases pursued but also by its intensity. One might expect the socially optimal level of enforcement to vary with economic activity as, for example, there are more merger notifications during booms and possibly more cartels during periods of weak demand. Furthermore, government preferences regarding the level and focus of enforcement may vary with the incumbent presidential administration.

The budgets of the DOJ and the Federal Trade Commission are indeed increasing in GDP (Kwoka 1999) but antitrust case activity is counter-cyclical (Ghosal and Gallo 2001). Although most studies do not find case activity to be related to the administration's political party, Ghosal (2004) shows that this is due to aggregation and mis-specification. He disaggregated data for 1958–2002 into criminal and civil cases and allowed there to be a structural break in the relationship between the usual independent variables – such as GDP, the DOJ's budget, and the president's political party – and the number of DOJ cases. Reasons for a break comprise the growing influence among economists and judges of the Chicago School – which argued that a number of previously considered antitrust offences may be profitable for firms to pursue for competitive reasons – and the fact that the Supreme Court had a two-thirds majority of Republican-nominated justices starting in 1972. Both of these forces would give less credence to certain practices – such as vertical restraints and monopolization practices – being treated as antitrust violations. A break in the number of civil cases (such as mergers and vertical restraints) occurred around the mid-1970s, which resulted in a significant decline, while a significant rise in the number of criminal cases (collusion) occurred around the late 1970s. There is also a post-regime rise in polarization between Republican and Democratic presidential administrations with Republicans pursuing more (less) criminal (civil) cases.

Impact of Antitrust Enforcement

Is enforcement having an effect? This is a difficult question for which hard facts are lacking, and sharply divergent views have been expressed. (See Baker 2003, and Crandall and Winston 2003; the latter should be read with caution as their review of some literatures is seriously deficient – Kwoka 2003, and Werden 2004, provide a critique.) With respect to the most egregious offence – namely, collusion – we pose three questions. Do cartels actually charge higher prices? Does prosecution lower prices? And, does successful prosecution have a deterrent effect?

The evidence is overwhelming that cartels raise prices. Connor and Lande (2005) have provided an exhaustive survey and found the median overcharge is 25 per cent. The evidence on how prices respond after indictment and conviction is mixed. A price decline was found in the break-up of cartels in white pan bread (Block et al. 1981);

and Feinberg (1984) found that, for four of five cartels, the Producer Price Index for the cartelized market fell by 6.6–11.4 per cent relative to a broader industry price index. Evidence to the contrary is provided in Sproul (1993) where, for 25 price-fixing cases over 1973–84, price (measured relative to that of a related good) *rose* by seven per cent in the four-year period after the indictment, although in some cases the immediate response was a nine to ten per cent fall in price. In light of the well-established evidence of an overcharge, the natural interpretation is that, although prosecution may reduce prices in the short run, in the longer run collusion may re-establish itself either explicitly or tacitly.

Even if prices do rebound from a conviction, prosecution and penalties are still useful because they reduce the profitability of collusion and thus may deter some cartels from forming. Indeed, there is some evidence of deterrence. The general method of testing for it is to have a reduced form equation explaining markups over time and to include a dummy variable when an action has been filed for collusion in a related market. In the case of white pan bread, markups fell for cities in a region for which the DOJ had filed an action that year in some other city in that region (Block et al. 1981). Similar evidence of deterrence holds for highway construction procurement auctions, which are notorious for bid-rigging (Block and Feinstein 1986).

In sum, the evidence is that cartels exist, they substantially raise price, and the indictment and conviction of firms may result in lower prices and may have a deterrent effect. Finally, financial penalties fall significantly short of making collusion unprofitable.

See Also

- ▶ Cartels
- ▶ Merger Analysis (United States)
- ▶ Merger Simulations

Acknowledgment I appreciate the comments of Vivek Ghosal.

Bibliography

Baker, J. 2003. The case for antitrust enforcement. *Journal of Economic Perspectives* 17(4): 27–50.

Block, M., and J. Feinstein. 1986. The spillover effect of antitrust enforcement. *The Review of Economics and Statistics* 68: 122–131.

Block, M., F. Nold, and J. Sidak. 1981. The deterrent effect of antitrust enforcement. *Journal of Political Economy* 89: 429–445.

Bryant, P., and E. Eckard. 1991. Price fixing: the probability of getting caught. *The Review of Economics and Statistics* 73: 531–536.

Connor, J. 2004. Effectiveness of antitrust sanctions on modern international cartels. Working paper. Purdue University.

Connor, J., and R. Lande. 2005. How high do cartels raise prices? Implications for optimal cartel fines. *Tulane Law Review* 80: 513–570.

Crandall, R., and C. Winston. 2003. Does antitrust policy improve consumer welfare? Assessing the evidence. *Journal of Economic Perspectives* 17(4): 3–26.

Feinberg, R.M. 1984. Strategic and deterrent pricing responses to antitrust investigations. *International Journal of Industrial Organization* 2: 75–84.

Gallo, J., K. Dau–Schmidt, J. Craycraft, and C. Parker. 2000. Department of Justice antitrust enforcement, 1955–1997: An empirical study. *Review of Industrial Organization* 17: 75–133.

Ghosal, V. 2004. Regime shifts in antitrust. Working paper. Georgia Institute of Technology.

Ghosal, V., and J. Gallo. 2001. The cyclical behavior of the Department of Justice's antitrust enforcement activity. *International Journal of Industrial Organization* 19: 27–54.

Harrington, J. Jr. 2005. Optimal corporate leniency programs. Working paper. Johns Hopkins University.

Harrington, J. Jr. 2006. Detecting cartels. In *Handbook of antitrust economics*, ed. P. Buccirossi. Cambridge, MA: MIT Press.

Kwoka, J. 1999. Commitment to competition: An assessment of antitrust agency budgets since 1970. *Review of Industrial Organization* 14: 295–302.

Kwoka, J. 2003. The attack on antitrust policy and consumer welfare: A response to Crandall and Winston. Working paper no. 03–008. Northeastern University.

Lande, R. 1993. Are antitrust 'treble' damages really single damages? *Ohio State Law Journal* 54: 115–174.

Motta, M. 2004. *Competition policy: Theory and practice*. Cambridge: Cambridge University Press.

Motta, M., and M. Polo. 2003. Leniency programs and cartel prosecution. *International Journal of Industrial Organization* 21: 347–379.

Page, W., eds. 1996. *Proving antitrust damages*. Chicago: American Bar Association.

Polinsky, A., and S. Shavell. 2000. The economic theory of public enforcement of law. *Journal of Economic Literature* 38: 45–76.

Spagnolo, G. 2003. *Divide et Impera*: Optimal deterrence mechanisms against cartels and organized crime. Working paper. University of Mannheim.
Sproul, M. 1993. Antitrust and prices. *Journal of Political Economy* 101: 741–754.
Viscusi, W., J. Harrington Jr., and J. Vernon. 2005. *Economics of regulation and antitrust*. 4th edn. Cambridge, MA: MIT Press.
Werden, G. 2004. Comment. *Journal of Economic Perspectives* 18(3): 224–225.
Werden, G., and M. Simon. 1987. Why price fixers should go to prison. *Antitrust Bulletin* 32: 917–937.

Anti-trust Policy

Oliver E. Williamson

Although many countries have adopted antitrust statutes and have an active antitrust enforcement programme, the United States was the first to enact national legislation on monopolies and monopolization. To be sure, English common law dealt with some of these matters long before the Sherman Act was passed in 1890. But the United States was and remains a leader in antitrust legislation, enforcement and research. The discussion herein focuses on the development of antitrust economics and related changes in antitrust enforcement within the United States.

Industrial Organization and, as a subfield therein, antitrust economics, is mainly a post-World War II development. The groundwork for this was laid by theoretical and empirical studies in the 1930s, of which Chamberlin's *Theory of Monopolistic Competition* (1933), Robinson's *The Economics of Imperfect Competition* (1933), Berle and Means's *The Modern Corporation and Private Property* (1932) and the series of studies by the Temporary National Economic Committee were especially important.

E.S. Mason was particularly influential in helping to give definition to the new field of Industrial Organization. Not only did he regard this as an important subject in its own right, but he perceived that informed antitrust enforcement was greatly in need of intellectual underpinnings. Interest in these matters mushroomed in the post-war period. Although the study of Industrial Organization was (and is) something of an art form, the leading texts – the one by Bain (1958), the other by Stigler (1968) – addressed the issues from the aforementioned standpoint of applied price theory. Not only was the firm regarded as a production function, but industry structure was thought to be virtually determinative of conduct and performance: 'an industry which does not have a competitive structure will not have competitive behavior' (Stigler 1952, p. 167). The structure–conduct–performance paradigm rapidly gained ascendancy.

The size distribution of firms (usually measured as a four-firm concentration ratio) and the condition of entry (usually assessed with reference to Bain's [1956] pioneering treatment of 'barriers to entry') were the key structural features. Non-standard or unfamiliar business conduct was believed to be suspect if not outright antisocial. A monopoly presumption was thus applied to vertical integration of activities on the periphery. It was widely and readily accepted that such a presumption applied *a fortiori* to nonstandard or unfamiliar contracting practices.

The past twenty years have witnessed a vast reshaping of antitrust economics. Antitrust enforcement and policy have followed these changes with a lag. Although market power remains a centrepiece, barriers to entry are now treated in a more discriminating way. Also, there is much greater respect for the benefits of economies than there once was. Nonstandard or unfamiliar business practices are no longer regarded as presumptively unlawful. And the study of strategic behaviour has emerged as a central antitrust economics and public policy concern. Consider these seriatim.

Barriers to Entry

Entry-barrier analysis made its appearance in the 1950s with the publication of books by Sylos-Labini (1956) and Bain (1956) and by Modigliani's formalization of the core argument (1958). It quickly made headway and was

virtually determinative of antitrust enforcement in the 1960s. Enforcement uses were sweeping, and successes came easily. Dissent nevertheless appeared as entry-barrier arguments came to be used uncritically.

Objections of two kinds were registered. For one thing, the entry-barrier models purportedly dealt with oligopoly without ever addressing how the mechanics of collective action were realized (Stigler 1968). Second, the existence of an 'entry barrier' was taken as a sufficient condition to warrant public-policy intervention. Comparative institutional analysis is not, however, concerned with defects judged with respect to a hypothetical ideal but with defects of a remediable kind.

Mistaken treatments of economies of scale are illustrative. To describe such a condition as a barrier to entry invites the conclusion that this is an antisocial outcome. Public policy hostility easily results. Thus the Federal Trade Commission declared that 'economic efficiency or any other social benefit [is] pertinent only insofar as it may tend to promote or retard the vigor of competition' (quoted in Bork 1978, p. 254), where competition is defined in structural terms. The Supreme Court evidently concurred. It thus flatly held in Procter & Gamble that 'possible economies cannot be used as a defense to illegality' (386 US 568, 574 [1967]).

This preoccupation with entry barriers predictably gave rise to perverse responses. Rather than ask for affirmative if not mitigating consideration, Procter & Gamble responded to the government's claims of economies in the Clorox case by first denying them and thereafter insisting that the government was unable definitively to prove that such economies existed. Such inverted reasoning could not and did not survive.

Economies as an Antitrust Defence

Although entry barrier analysis made rudimentary use of price theory, it made little appeal to applied welfare economics. This is regrettable, since application of the basic partial equilibrium welfare economics model to an assessment of the allocative efficiency trade-off between market power and economies disclosed that to sacrifice economies in favour of smaller price–cost margins often came at a high cost (Williamson 1968b). Albeit subject to qualification, this view has made progressive headway (Liebeler 1978; Bork 1978; Muris 1979; Fisher and Lande 1983). Indeed, the 1984 Merger Guidelines of the Department of Justice now invite firms proposing a merger to present evidence of efficiencies. Although this is not without enforcement hazards, antitrust enforcement excesses of the 1960s – which led to the suppression, denial and perverse interpretation of efficiency – have been generally discredited.

Public policy towards vertical mergers has been similarly transformed. Thus, whereas the 1968 Vertical Merger Guidelines employed firm-as-production-function reasoning, whence vertical integration was proscribed if there was 'an appreciable degree of market control' at any stage in the system (Stigler 1955, p. 183), the earlier limits have been relaxed under the firm-as-governance-structure approach to the study of economic organization (Coase 1937; Williamson 1985). Not only do the current Merger Guidelines make express provision for transaction cost economies, but they acknowledge that the characteristics of investments (especially the condition of asset specificity) are germane to an assessment of economic benefits. Vertical integration is now held to be problematic only where the market structure would support strategic behaviour.

Vertical Market Restrictions

The view that vertical market restrictions are presumptively anti-competitive has likewise been abandoned. The 'inhospitality tradition' maintained that nonstandard contracting practices (tie-ins, block booking, customer and territorial restrictions, and the like) had the purpose and effect of realizing leverage, facilitating price discrimination or erecting barriers to entry. More recent scholarship has taken a broader view of these matters. Two factors have been responsible for the new learning.

For one thing, the technological dichotomy between firm and market organization has been supplanted by a transactional view in which the existence and economic merits of a wide range of intermediate ownership and contracting modes are admitted. As a consequence, the earlier monopoly presumption has given way to an examination of transactions and the costs that attend alternative forms of contracting. Second, the focus on the *ex post* effects of contractual restraints has been supplanted by a more complete treatment of contract in which the *ex ante* bargain and the *ex post* terms are regarded simultaneously. The economic importance of intertemporal contractual integrity is emphasized by this latter perspective, whereas the focus of earlier piecemeal interpretations of contract had been on the momentary relationship between the parties. Different assessments of, for example, franchise terminations often result.

Thus, whereas the piecemeal approach to contract focuses on power disparities between the franchisor and the franchisee, the intertemporal approach examines *ex post* behaviour in relation to the *ex ante* bargain and asks whether efficiency considerations and reputation effects are operative. Successive studies of vertical restraints from an intertemporal perspective (Telser 1960, 1981; Williamson 1979, 1983; Klein and Leffler 1981) urge that vertical market restrictions ought not to be regarded as presumptively unlawful but that their anticompetitive effects should be judged in strategic behaviour terms.

Inasmuch as conglomerate organization is at best an anomaly within the firm-as-production-function framework, conglomerate mergers in the pre-1970 period operated under a cloud. The monopoly presumption that was ascribed to nonstandard practices was thought to apply, whence vague monopoly purpose was imputed to conglomerate mergers. The 1968 Merger Guidelines of the Department of Justice, for example, held that 'Since reciprocal buying ... is an economically unjustified business practice which confers a competitive advantage on the favored firm unrelated to the merits', conglomerate mergers which create a prospect of reciprocity will ordinarily be challenged.

Subsequent study of nonstandard contracting, however, revealed that reciprocal trade can also serve efficiency purposes. In particular, reciprocity (of an appropriate kind) can help to create a mutual 'credible commitment'. To be sure, only the subset of contracts where trade is supported by investments in transaction specific assets will warrant such an efficiency rationale. Earlier claims that reciprocity is meritless, however, cannot be sustained.

Out of awareness, presumably, that the monopoly presumption had been overdone, the 1982 and 1984 Merger Guidelines are silent with respect to the special dangers of reciprocity. Instead, conglomerate acquisitions, which is the context where the reciprocity issue was originally expressed, are now held to pose antitrust problems only if the condition of potential entry is adversely effected. This is a much narrower conception and is one with which antitrust is legitimately concerned.

The upshot is that antitrust theory and policy were vastly transformed over the interval 1965–1985. Policy changes have followed theory, with lags of ten years and more as new theory was first subjected to the crucible of academic discourse. Although there is a real possibility that the antitrust pendulum could swing too far, the conceptual errors of the 1960s are unlikely to be repeated.

Strategic Behaviour

Antitrust is in no position to settle for the quiet life. Issues of strategic business behaviour exploded onto the antitrust scene in the 1970s and have been prominently featured on the research and enforcement agenda since. A series of prominent antitrust suits and growing academic interest in business strategy were jointly responsible.

Major antitrust suits alleging predation were brought both by private firms and the Federal Trade Commission. Several private antitrust suits alleging predatory pricing by IBM against computer peripheral manufacturers illustrate the former. The Federal Trade Commission advanced

novel theories of strategic anti-competitive behaviour in asserting predatory brand proliferation in the ready-to-eat cereals industry and pre-emptive investments in the titanium dioxide industry.

Academic interest in predation and, more generally, in strategic business behaviour was both responsive to and contributed to these antitrust developments. A consensus has yet to be reached, however, on such basic matters as the appropriate criteria for judging price predation.

Lack of such agreement encourages some to argue that predatory pricing is an antitrust fiction and ought to be disregarded. This derives, however, from a static assessment of predation. But the nub of the problem is intertemporal and features uncertainty. The study of these matters remains in flux.

A separate but nevertheless related academic literature has reformulated the earlier entry barrier work on more secure economic foundations. The study of pre-emptive investments (Spence 1977) was successively elaborated by introducing the concept of credible threat (Dixit 1979, 1980; Eaton and Lipsey 1980, 1981). As it turned out, an investment took on credibility in the degree to which it was non-redeployable – which is precisely the issue with which the asset specificity literature is concerned. Subsequent work has extended the study of strategic behaviour to include an assessment of innovation (von Weizsäcker 1980; Ordover and Willig 1981) and to introduce probablistic gaming considerations into the calculus of predation (Milgrom and Roberts 1982; Kreps and Wilson 1982).

The study of strategic behaviour in the context of 'raising rivals costs' has also been progressing. The use of wage rates as a barrier to entry and of strategic forward integration (Williamson 1968a, 1979) has since been generalized to encompass a wide class of cost-increasing strategies (Salop and Scheffman 1983).

A third factor contributing to concern over and interest in the study of strategic behaviour has been posed by international competition. Allegations that strategic business behaviour – with respect both to foreign and domestic markets – is sometimes aided and abetted by government agencies are widespread. The joinder of the Industrial Organization literature and of the International Trade literature is needed to develop these issues in a more rigorous and systematic way. Work of this kind is in progress and is nicely summarized in Grossman and Richardson (1985). That there are very real and serious problems posed for which we do not presently have well-defined answers is plainly the case. Caution against protectionist use (or abuse) of antitrust to insulate markets against legitimate international competition is a matter of real concern. But the proposition that strategic behaviour is a myth in this and other contexts is simplistic and repeatedly refuted by the facts. Unpacking these issues poses a major intellectual challenge in the years ahead.

See Also

- ▶ Cartels
- ▶ Concentration Ratios
- ▶ Monopoly
- ▶ Regulation and Deregulation

References

Bain, J. 1956. *Barriers to new competition*. Cambridge, MA: Harvard University Press.

Bain, J. 1959. *Industrial organization*. New York: Wiley.

Berle, A.A., and G.C. Means. 1932. *The modern corporation and private property*. New York: Macmillan.

Bork, R.H. 1978. *The antitrust paradox*. New York: Basic Books.

Chamberlin, E. 1933. *Theory of monopolistic competition*. Cambridge, MA: Harvard University Press.

Coase, R.H. 1937. The nature of the firm. *Economica* 4: 386–405. Reprinted in *Readings in price theory,* ed. G.-J. Stigler and K.E. Boulding. Homewood: Richard D. Irwin, 1952.

Dixit, A. 1979. A model of duopoly suggesting a theory of entry barriers. *Bell Journal of Economics* 10: 20–32.

Dixit, A. 1980. The role of investment in entry deterrence. *Economic Journal* 90: 95–106.

Eaton, B.C., and R.G. Lipsey. 1980. Exit barriers are entry barriers: The durability of capital. *Bell Journal of Economics* 11: 721–729.

Eaton, B.C., and R.G. Lipsey. 1981. Capital commitment and entry equilibrium. *Bell Journal of Economics* 12: 593–604.

Fisher, A., and R. Lande. 1983. Efficiency considerations in merger enforcement. *California Law Review* 71: 1580–1696.

Grossman, G.M. and Richardson, J.D. 1985. Strategic trade policy: A survey of issues and early analysis. Special Papers in International Economics No. 15, Princeton University.
Klein, B., and K.B. Leffler. 1981. The role of market forces in assuring contractual performance. *Journal of Political Economy* 89: 615–641.
Kreps, D.M., and R. Wilson. 1982. Reputation and imperfect information. *Journal of Economic Theory* 27: 253–279.
Liebeler, W.C. 1978. Market power and competitive superiority in concentrated industries. *UCLA Law Review* 25: 1231–1300.
Mason, E. 1957. *Economic concentration and the monopoly problem*. Cambridge, MA: Harvard University Press.
Milgrom, P., and J. Roberts. 1982. Predation, reputation, and entry deterrence. *Journal of Economic Theory* 27: 280–312.
Modigliani, F. 1958. New developments on the oligopoly front. *Journal of Political Economy* 66: 215–232.
Muris, T.J. 1979. The efficiency defense under section 7 of the Clayton Act. *Case Western Reserve Law Review* 30: 381–432.
Ordover, J.A., and R.D. Willig. 1981. An economic definition of predatory product innovation. In *Strategic views of predation*, ed. S. Salop, 301–396. Washington, DC: Federal Trade Commission.
Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.
Salop, S., and D. Scheffman. 1983. Raising rivals' costs. *American Economic Review* 73: 267–271.
Spence, A.M. 1977. Entry, capacity investment and oligopolistic pricing. *Bell Journal of Economics* 8: 534–544.
Stigler, G.J. 1952. The case against big business. *Fortune* 47, May, 123 et seq.
Stigler, G.J. 1955. Mergers and preventive antitrust policy. *University of Pennsylvania Law Review* 104: 176–184.
Stigler, G.J. 1968. *The organization of industry.* Homewood: Richard D. Irwin.
Sylos-Labini, P. 1956. *Oligopoly and Technical Progress.* Trans. Elizabeth Henderson. Cambridge, MA: Harvard University Press, 1962.
Telser, L. 1960. Why should manufacturers want fair trade? *Journal of Law and Economics* 3: 86–104.
Telser, L. 1981. A theory of self-enforcing agreements. *Journal of Business* 53: 27–44.
von Weizsäcker, C.C. 1980. *Barriers to entry.* New York: Springer.
Williamson, O.E. 1968a. Wage rates as a barrier to entry: The Pennington case in perspective. *Quarterly Journal of Economics* 82: 85–116.
Williamson, O.E. 1968b. Economies as an antitrust defense: The welfare tradeoffs. *American Economic Review* 58: 18–35.
Williamson, O.E. 1979. Assessing vertical market restrictions. *University of Pennsylvania Law Review* 127: 953–993.
Williamson, O.E. 1983. Credible commitments: Using hostages to support exchange. *American Economic Review* 73: 519–540.
Williamson, O.E. 1985. *The economic institutions of capitalism*. New York: Free Press.

Antonelli, Giovanni Battista (1858–1944)

A. P. Kirman

Keywords
Antonelli, G. B.; Engel curves; Indirect demand function; Integrability problem

JEL Classifications
B31

Antonelli was born near Pisa in 1858. He studied mathematics and then went on to qualify as an engineer. Although his life was devoted to civil engineering, he made an important contribution to early mathematical economics. His *Sulla teoria matematica dell'economia politica* (1886), intended to be the first part of a book, is remarkable, in particular for the conditions he gives for the 'integrability problem'.

This asks under what conditions single valued demand functions are generated by the maximization of a utility function. Antonelli studied the 'local' aspects of this problem. He started from what is now called the indirect demand function:

$$p = M[q]$$

where q is the vector of goods and p the vector of prices. He gave the symmetry of the matrix of the price substitution terms $\partial p_i/\partial q_j$ as a condition for the recoverability of the utility function but should have also required the negative semi-definiteness of this matrix. The importance of this work has been recognized by Samuelson (1950) and later authors, but passed unappreciated if not unnoticed at the time.

In the same work Antonelli derives a condition for a market demand function to be derivable from a market utility function, that is, that individuals have linear parallel Engel curves. This condition was found much later by Gorman (1953) and Eisenberg (1961). Antonelli had an active and productive career in engineering and what would now be called 'operations research' but never came back to theoretical economics. He died in 1944.

Selected Works

1886. *Sulla teoria matematica dell'economia politica.* Pisa. Reprinted, with an introduction by G. Demaria, Milan: Malfasi, 1952.

Bibliography

A detailed and careful account of Antonelli's contributions and a translation of his economic paper is given in:

Chipman, J.S., Hurwicz, L., Richter, M.K. and Sonnenschein, H.F., eds. 1971. *Preferences*, *utility and demand.* New York: Harcourt Brace and Jovanovich. (See in particular Introduction to Part II by J.S. Chipman, chapter 16: a translation of Antonelli and chapter 9 on the integrability problem by L. Hurwicz.)

Eisenberg, E. 1961. Aggregation of utility functions. *Management Science* 7: 337–350.

Gorman, W.M. 1953. Community preference fields. *Econometrica* 21: 63–80.

Samuelson, P.A. 1950. The problem of integrability in utility theory. *Economica* 17: 355–385.

Aoyama, Hideo (born 1910)

Mitsuo Saito

Aoyama was born in Okayama, Japan. He obtained an MA (1932) and a doctorate (1951) in economics at Kyoto University, where he was Professor of Economics from 1946 to 1973.

In 1937 Aoyama wrote 'The Economic Theory of Monopoly', the first monograph on mathematical economics written in Japanese. He constructed a systematic classification of markets, providing a mathematical model of price determination for each type of market. His theory traces a path from isolated exchanges with numerous equilibrium points on a segment of the Edgeworth contract curve, towards perfect competition, where the equilibrium must be located on a particular point on the same curve. This view resembles the more recent theory of the core of an economy.

Between 1938 and 1943 Aoyama published a number of articles on the dynamics of economic fluctuation; these are included in the three volumes of his Collected Papers (1949, 1950 and 1953). He stressed the significance of distinguishing between the theory of the temporary state (in modern terminology, temporary general equilibrium) and the theory of process over time (macrodynamics). His general equilibrium model of production (1938) – influenced by the work of F.H. Knight, Gunnar Myrdal and Ragnar Frisch – expressed the supply and demand of all commodities as a function of their present prices, expected prices, and stocks. In the light of Hicks's *Value and Capital* (1939) he provided a rigorous mathematical treatment of the concept of the composite commodity grouping (1943). He was much influenced by the work of D.H. Robertson and attempted to reformulate his period analysis (1941a, b).

Another of his interests was the theory of crisis, particularly Spiethoff's and Tugan-Baronovsky's interpretation of overinvestment. He demonstrated that Say's Law was an essential element in Spiethoff's theory and that the theory could not stand because errors in expectation were bound to cause general overproduction.

Like Yasuma Takata, his predecessor in the Kyoto chair of economics, Aoyama was a sociologist as well as an economist. In the 1940s he became particularly interested in Max Weber and in his main works of this period (1948, 1949) he developed a typology of national economies based on Weber's theory of ideal types. He emphasized the common characteristics of rational systems of control such as the military, the bureaucracy and the corporation.

Aoyama was a pioneer of the scientific analysis of the economy and society during the dark age of Japanese social science. He introduced current

Western theories to his students and influenced the new generation of Japanese mathematical economists in the postwar era.

Selected Works

1941a. A critical note on D.H. Robertson's theory of savings and investment (I). *Kyoto University Economic Review* 16(1): 49–73.
1941b. A critical note on D.H. Robertson's theory of savings and investment (II). *Kyoto University Economic Review* 16(2): 64–81.
1943. On the extension of the concept of a commodity: a note on Hicks' theory of the 'group of commodities' *Kyoto University Economic Review* 18(2): 48–68.
1944. Die Rechnungsmässige Rationalität als Grundlegendes Merkmal der Modernen Volkswirtschaft, I. *Kyoto University Economic Review* 19(1): 44–60.
1959. (With Toru Nishikawa.) Business fluctuations in the Japanese economy in the inter-war period. *Kyoto University Economic Review* 28(1): 14–39.

Appropriate Technology

Alice H. Amsden

Depending on the decade, different criteria have provided the basis for judging the appropriateness of technology in developing countries. In the 1950s and 1960s, debate centred on whether the choice of technique ought to be guided by the objective of maximizing the growth rate, rather than the level, of output. In the 1970s, the maximand became employment. There was a surge in articles on employment creation through income redistribution and on the merits, more in theory than in practice, of alternative technology life-styles. Soon after, what was considered to be appropriate came once again to mean a competitive market outcome. So defined, the term 'appropriate technology' served no distinct purpose and dropped out of use.

Central to the economic literature in the 1950s and 1960s on how to accelerate development was lengthy discussion of choice of technique. A socialist like Dobb (1963) argued that insofar as the limiting factor consists in the available surplus of foodstuffs and other consumer goods, it will not be the best policy (from the growth standpoint) to invest in low productivity, 'labour-intensive' techniques (as advocated by the twin doctrines of Marginal Productivity and Comparative Cost). Techniques should be chosen that, by achieving a higher level of output per worker, make the surplus product larger. Dobb's point was also made by some adherents of the neoclassical production function who faulted the market model for being too preoccupied with static resource allocation (Sutcliffe 1974, ch. 5).

Output grew very rapidly in developing countries in this period. Yet outside the socialist bloc, employment growth stagnated. The term 'appropriate technology' was popularized by economists seeking to understand this fact and what could be done about it.

The reasons proffered for slow employment growth amidst rapid rises in output were wide ranging. The majority blamed market distortions. The political clout of urban workers raised wages, subsidization of credit cheapened capital, and overvalued exchange rates invited machinery imports. Another group offered managerial explanations. A choice of technique was not a mere theoretical abstraction, as evidenced by different factor proportions among plants of the same firm ostensibly producing the same product in different countries. But 'engineering man' rather than *homo economicus* did the choosing, and in developing countries chose production processes to raise technical sophistication, reduce labour problems and enhance product quality (Stobaugh and Wells 1984). A third set of reasons was elaborated by Frances Stewart (1972, 1985). First, she recognized a maximization conflict not merely between output and the growth of output, but also between output and employment. Two techniques of different factor proportions might incur the same total costs, but the technique with a

higher ratio of capital to labour might produce higher output. In theory, this technique might generate more employment in the long run, but over time, new technologies would be devised, even more capital-using than those displaced, because innovation occurred in high wage countries. The assumption of a continuous spectrum of techniques to produce a given output, therefore, was belied by history (so lowering wages might succeed only in raising rents rather than employment). At the other extreme, 'technological determinists' were equally wrong in contending the absence of any choice. Yet by varying operating scale and discriminating among products to satisfy consumer wants, the technical menu could be widened.

By this line of reasoning, increasing employment was seen to depend upon choosing appropriate *products* and stimulating small-scale industry. The possibility of specializing through foreign trade in labour-intensive manufactures was underplayed while it was assumed that the consumption bundle of the poor generally involved more labour-intensive production processes than that of the rich. Therefore, employment would rise with a redistribution of income from rich to poor to finance 'informal sector' markets.

The most radical articulation of this view, in the Proudhonian sense, was heard on the fringes of the economics profession. Adherents were sufficiently focused and active to be called a movement, the AT. The Appropriate Technology movement traced its intellectual heritage to Mahatma Gandhi and E.F. Schumacher (1973) and emphasized rural community development. Appropriate technology came to embody self-reliance, a rejection of the technico-economic values of industrialized nations, the use of locally available resources, especially solar energy, and not just a higher ratio of labour to capital (Jequier 1976). Appropriate technology had to be developed by and for the people who lived by it.

Politically, even the most radical tendencies in the AT movement failed to mobilize mass support. In policy making, not even the moderates made much headway, except possibly in India. Political economy constraints had largely been ignored and demands for income redistribution proved fanciful. The difficulties of devising new technologies and developing local capabilities to assimilate and improve them were underestimated. And at best, all that might be expected was a small surplus and slow growth.

Into this void stepped enthusiasts of export-led growth. Through trade, developing countries could specialize in 'labour-intensive' products without having to redistribute income. With a reduction in market distortions, the choice by profit-maximizing firms of appropriate technology would be automatic. But the conflict between output growth and employment inherent in this reading of appropriate technology was never well understood. Governments in middle income countries with a long-term perspective, appreciative of the high risks of a trade reliant development strategy amidst rising wages, borrowed heavily abroad. Funds were used to establish new industries with long lead times and high capital requirements in order to stay ahead of competition from even lower wage countries in 'labour-intensive' goods. Debt servicing required access to the markets of advanced countries. Yet as advanced countries climbed up the ladder of comparative advantage, they became less capable of achieving full employment and less willing to relinquish their labour-using industries to imports. Neither the theoretical solution, lower wages, nor the politically popular one, protection, promised relief for indebted nations. Thus, the technologies that seemed appropriate to different groups of countries were out of synchronization. The fast growing, newly industrializing countries and the slower growing advanced economies collided in a widening range of markets for industrial products.

See Also

- ▶ Backwardness
- ▶ Choice of Technique and the Rate of Profit
- ▶ Schumacher, E.F. (Fritz) (1911–1977)

Bibliography

Dobb, M. 1963. *Economic growth and underdeveloped countries*. London: Lawrence & Wishart.

Jequier, N. 1976. *Appropriate technology*. Paris: Development Centre of the Organization for Economic Cooperation and Development.
Schumacher, E.F. 1973. *Small is beautiful*. London: Blond & Briggs.
Stewart, F. 1972. Choice of technique in developing countries. *Journal of Development Studies* 9(1): 99–121.
Stewart, F. 1985. Macro policies for appropriate technology: An introductory classification. In *Technology, institutions and government policies*, ed. J. James and S. Watanabe. London: Macmillan.
Stobaugh, R., and L.T. Wells Jr. (eds.). 1984. *Technology crossing borders*. Boston: Harvard Business School Press.
Sutcliffe, R.B. 1971. *Industry and underdevelopment*. London: Addison-Wesley.

Approval Voting

Enriqueta Aragones and Micael Castanheira

Abstract

In a single-winner voting system, approval voting gives voters the possibility to cast a ballot for (or 'approve of') as many candidates as they wish – that is, voters are freed from the constraint of voting for only one candidate. The candidate receiving the greatest total number of votes is declared the winner. Approval voting has several compelling advantages over other voting procedures, and has been used by various governments and organisations around the world.

Keywords

Approval voting; Condorcet winner; Strategic voting; Voting systems

JEL Classifications

D71; D72

Robert J. Weber coined the term 'approval voting' to describe an election system in which each voter is allowed to vote for as many candidates as they wish – that is, voters can 'approve of' all the candidates deemed 'acceptable', and the candidate receiving the greatest total number of votes is declared the winner.

Scholarly analyses of this voting system began in the 1970s, with the works by Steven Brams, Peter Fishburn and Robert J. Weber (Brams and Fishburn 1978, 1983, 2007; Weber 1995), and led to an outburst of research that is still ongoing today. Their germinal motivation lies in some of the weaknesses of the plurality (or first-past-the-post) voting system, in which voters can vote for only one candidate and the candidate with the most votes wins.

Issues with Plurality Voting

We can readily identify three issues with plurality voting. First, a group of people may be a minority and yet represent a plurality. Minority candidates who would lose a one-to-one electoral contest may thus win the election. Second, to prevent such an outcome, voters may need to adopt an insincere strategy, and concentrate their ballots on a strong contender instead of their preferred candidate. (In this system, voting is 'sincere' if voters cast their ballot for their preferred candidate.) Strategic mistakes may then lead to 'wrong' electoral outcomes: voters may fail to coordinate on the right candidate. Third, candidates have an incentive to design their platform so as to be very strong in some subgroups of the population, instead of trying to reach wider consensus.

Some Properties of Approval Voting

Under approval voting (AV), a 'sincere' voting strategy can be summarised by the 'worst' candidate that the voter wants to approve of. That is, each voter decides of a cutoff that divides acceptable candidates from unacceptable ones. All the candidates above the cutoff are then approved of.

A common concern is that voters who are almost indifferent between several candidates may find many of them acceptable. However, they cannot exert more voting power than someone who finds only one candidate acceptable. First, each voter can only cast one single ballot for a

given candidate. Second, voting for many candidates actually dilutes the voter's ballot: approving of all candidates is equivalent to abstaining.

Brams and Fishburn (2005) highlight six important advantages of AV:

- It gives voters more flexible options: beyond what they can do under plurality, voters can also vote for additional candidates.
- It helps elect the strongest candidate: candidates who attract a broad consensus will be approved of by more voters.
- It reduces negative campaigning: candidates have an incentive to broaden their appeal to reach for the approval of voters who have a different first choice.
- It increases voter turnout: being better able to express their preferences, voters are more likely to vote.
- It will give minority candidates their proper due. Minority candidates receive their true level of support under AV: supporters of a minority candidate need not abandon their preferred candidate to lend support to stronger candidates.
- It is eminently practicable: it is simple for voters to understand and use. (Several experiments were run to verify that voters do indeed understand how to behave under such an electoral system. See for instance Laslier and Van der Straeten 2008.) It can also readily be implemented on existing voting machines.

Controversies

The initial perception that AV produces sharp predictions has been questioned by subsequent research. Saari and van Newenhizen (1988) show that the multiplicity of 'sincere' strategies generates a problem of outcome indeterminacy. Niemi (1984) shows that it 'almost begs voters to behave strategically'. Using the 'natural experiment' of the US Electoral College in 1800, Nagel (2007) argues that these strategic considerations may often produce tied outcomes. Yet Myerson and Weber (1993) show that Condorcet winners must always be among the likely winners.

Applications

While scholarly analyses of AV started in the 1970s, electoral systems in which voters can cast either a single or a multiple ballot had been used previously. For instance, the US Electoral College, used AV between 1788 and 1800 (Nagel 2007). In a more distant past, related rules have been used in Venice (Lines 1986) and in papal elections (Colomer and McLean 1998). In those instances, AV was not maintained, partly because the technology to count votes was still primitive. Counting more than one vote per elector was thus costly. In addition, the size of the electorate was relatively small, which in AV heightens the incentives of strategic manipulation. The fruits of AV can indeed be expected to be most ripe when the size of the electorate is sufficiently large: strategic manipulations by one voter are then less likely to influence the outcome.

The outburst of academic research and the proven desirable properties of AV led several scientific societies to adopt it for their internal elections (Brams and Fishburn 2005). In the former Soviet Union, many elections involved a similar system. It is also used to organise referenda in some US states, and to elect the secretary-general of the United Nations.

See Also

▶ Strategic Voting
▶ Voting Paradoxes

Bibliography

Brams, S.J., and P.C. Fishburn. 1978. Approval voting. *American Political Science Review* 72: 831–847.

Brams, S.J., and P.C. Fishburn. 1983. *Approval voting.* (2nd ed. 2007). New York: Springer.

Brams, S.J., and P.C. Fishburn. 2005. Going from theory to practice: The mixed success of approval voting. *Social Choice and Welfare* 25: 457–474.

Colomer, J., and I. McLean. 1998. Electing popes: Approval balloting and qualified majority rule. *Journal of Interdisciplinary History* 29: 1–22.

Laslier, J.-F., and K. Van der Straeten. 2008. A live experiment on approval voting. *Experimental Economics* 11: 97–105.

Lines, M. 1986. Approval voting and strategic analysis: A Venetian example. *Theory and Decision* 20: 155–172.
Nagel, J. 2007. The Burr dilemma in approval voting. *Journal of Politics* 69: 43–58.
Niemi, R.G. 1984. The problem of strategic behavior under approval voting. *American Political Science Review* 78: 952–958.
Myerson, R., and R. Weber. 1993. A theory of voting equilibria. *American Political Science Review* 87: 102–114.
Saari, D.G., and J. Van Newenhizen. 1988. The problem of indeterminacy in approval, multiple, and truncated voting systems. *Public Choice* 59: 101–120.
Weber, R.J. 1995. Approval voting. *Journal of Economic Perspectives* 9: 39–49.

Approximate Solutions to Dynamic Models (Linear Methods)

Harald Uhlig

Abstract

This article explains how to obtain an approximate solution to dynamic stochastic discrete-time (DSGE) models by first log-linearizing the relevant equations and then obtaining a recursive law of motion, by using the method of undetermined coefficients. Calculations are provided based on both an eigenvector decomposition and the QZ or Schur decomposition. The role of sunspots and the relationship to the method of Blanchard and Kahn are discussed. The base example is a generic real business cycle model, for which log-linearization is described generally and in detail. The method described should be easily implemented. Further literature references and software sources are provided.

Keywords

Approximate solutions to dynamic models (linear methods); Dynamic stochastic general equilibrium (DSGE) models; Linearization; Log-linearization; QZ decomposition; Real business cycles; Representative agent; Recursive law of motion; Sunspots

JEL Classifications

D4; D10

Linear methods are often used to compute approximate solutions to dynamic models, as these models often cannot be solved analytically. While a plethora of advanced numerical methods exist, the most popular 'bread-and-butter' method for solving them is linearization. It is described here first with the example of a simple real business cycle model, but is applicable generally to dynamic stochastic general equilibrium (DSGE) models. It is shown how to easily generate the log-linearized equations needed. The linear system is then solved for the recursive law of motion, by using the method of undetermined coefficients. The classic reference for solving linear difference models under rational expectations is Blanchard and Kahn (1980), while Kydland and Prescott (1982) is the origin of the modern approach of calculating numerically approximate solutions to dynamic stochastic models in order to obtain quantitative results. Much of the material here is taken from Uhlig (1999), which builds on the method of undetermined coefficients in King et al. (2002).

A Basic Example

As a basic example, consider a version of the real business cycle model of Hansen (1985). A social planner or representative agent chooses c_t, k_t, y_t, l_t and n_t to maximize the utility function $U = E\left[\sum_{t=0}^{\infty} \beta^t u(c_t, l_t)\right]$ for some twice differentiable utility function u(.), satisfying the usual conditions, subject to the constraints

$$c_t + k_t = y_t + (1-\delta)k_{t-1} y_t = \gamma_t f(k_{t-1}, n_t) 1 = n_t + l_t$$

as well as a given initial capital stock k_{-1},where c_t denotes consumption, k_t denotes capital, y_t denotes output, l_t denotes leisure, n_t denotes labour, $f(k,n)$ denotes a twice differentiable production function, typically assumed to obey

constant returns to scale, β is the discount factor and γ_t is total factor productivity, with $z_t = \log(\gamma_t) - \log(\gamma^*)$ evolving according to $z_t = \rho z_{t-1} + \varepsilon_t$ where $E_t[\varepsilon_{t+1}] = 0$ for some values γ^* and ρ, with $-1 < \rho < 1$. A solution is a stochastic sequence $(c_t, k_t, y_t, l_t, n_t)$, $t \geq 0$ where all variables dated t are independent of all ε_s for $s > t$ and satisfies all constraints, and which maximizes the utility function given above within the set of all such sequences.

The necessary first-order conditions for this problem are given by

$$\begin{aligned} u_c(c_t, l_t) &= \lambda_t u_l(c_t, l_t) = f_n(k_{t-1}, n_t)\lambda_t \\ &= \beta E_t[\lambda_{t+1} R_{t+1}] R_t \\ &= f_k(k_{t-1}, n_t) + 1 - \delta \end{aligned}$$

Linearization

The first step towards solving the model by linear approximation is to linearize all the constraints and necessary equations (possibly after substituting out some variables, if so desired). Linearization amounts to finding a first-order approximation to all equations. Formally, linearization amounts to replacing a set of equations $0 = g(x_t)$ in a vector x_t of variables with its linearized counterpart around some point of approximation x^*, $0 = g(x^*) + g'(x^*)\tilde{x}_t$ where $\tilde{x}_t = x_t - x^*$ is the deviation of x_t from the approximation point x^* and where $G'(x^*)$ is the matrix of first derivatives of $G(.)$. As point of approximation x^*, the nonstochastic steady state is often chosen, that is, one solves the equations $0 = g(x^*)$ under the assumption that all exogenous stochastic variables are constant (here: $\gamma_t = \gamma^*$ and all $\varepsilon_s = 0$). Then, the remaining linearized system consists of $0 = g'(x^*)\tilde{x}_t$.

Since many economic variables are constrained to be positive, it is often more attractive to log-linearize the equations rather than to linearize them. The difference between linearization and log-linearization is that entries in x_t denote the original variable (for example, consumption c_t) in the case of linearization and the log of these variables (for example, $\log(c_t)$) in the case of log-linearization. There is no need to choose either linearization or log-linearization for all entries in x_t. One may choose to linearize some and log-linearize others or take other transformations. Indeed, for variables such as trade balances it is better to use linearization rather than log-linearization, if they can take negative values. Also, tax rates, for example, are often more appropriately linearized than log-linearized to provide a more useful interpretation.

This makes no difference as far as the linearized solution is concerned. More generally, differentiable and differentiable invertible transformations (that is, homeomorphisms) of the variables (for example, taking ratios of variables) make no difference to the properties of the linearized solution. The differences always lie only in the recalculation of the original variables, where one may want to take into account the nonlinearities originally inherent in the model. To see, more generally, that any homeomorphism (that is, differentiable and differentiably invertible transformation) $y_t = h(x_t)$ of the variables makes no difference to remaining calculations, note that the equations can be restated as $0 = g(h^{-1}(y_t))$. The linearized version is now $0 = g\left(h^{-1}(y^*)\right) + g'(x^*)\left(f^{-1}\right)'(y^*)\tilde{y}_t$, which coincides with the previous linearization if $y^* = F(x^*)$, noting that $\hat{y} = f'(y^*)\hat{x}_t$ as well as $I = f'(x^*)(f-1)'(y^*)$.

While linearization can be performed numerically or with the usual rules of calculus, one can often 'read' the log-linearized version of an equation from its original form, exploiting $x_t = \exp(y_t) \approx x^* + x^*\tilde{y}_t$, where now $y_t = \log(x_t)$. Write $\hat{x}_t$ instead of $\tilde{y}_t$ for the loglinear deviation.

For log-linearization, the following useful 'rules' can easily be derived. Let a_t, b_t, c_t be three variables, with $c_t = h(a_t)$ for some monotone and differentiable function h(.), and let B be some constant. Then,

$$\begin{aligned} a_t + Bb_t &\approx (a^* + Bb^*) + \left(a^*\hat{a}_t + Bb^*\hat{b}_t\right) Ba_t b_t \\ &\approx (Ba^*b^*) + (Ba^*b^*)\left(\hat{a}_t + \hat{b}_t\right)\hat{c}_t \\ &\approx \frac{h'(a^*)a^*}{h(a^*)}\hat{a}_t \end{aligned}$$

Either with these rules or directly, the equations in the example log-linearize to

$$
\begin{aligned}
c^*\hat{c}_t k^*\hat{k}_t &= y^*\hat{y}_t + (1-\delta)k^*\hat{k}_{t-1}\hat{y}_t \\
&= -z_t + \frac{f_k k^*}{f}\hat{k}_{t-1} + \frac{f_n n^*}{f}\hat{n}_t 0 \\
&= n^*\hat{n}_t + (1-n^*)\hat{l}_t\hat{\lambda}_t \\
&= \frac{u_{cc}c^*}{u_c}\hat{c}_t + \frac{u_{cl}l^*}{u_c}\hat{l}_t \frac{u_{cl}l^*}{u_l}\hat{c}_t \frac{u_{ll}l^*}{u_l}\hat{l} \\
&= \frac{f_{nk}k^*}{f_n}\hat{k}_{t-1} + \frac{f_{nn}n^*}{f_n}\hat{n}_t\hat{\lambda}_t \\
&= E_t\left[\hat{\lambda}_{t+1} + \hat{R}_{t+1}\right]R^*\hat{R}_t \\
&= \frac{f_{kk}k^*}{f_k}\hat{k}_{t-1}\frac{f_{kn}k^*}{f_k}\hat{n}_t.
\end{aligned}
$$

Solving for the Recursive Law of Motion

With some further algebra, one can turn this system into a second-order one-dimensional difference equation, $0 = E_t[Fx_{t+1} + Lz_{t+1}] + Gx_t + Mz_t + Hx_{t-1}$ plus the evolution of the exogenous state, $z_t = Nz_{t-1} + O\varepsilon_t$, where $x_t = k_t$ is the capital stock, and F, L, G, M, H, N and O are real numbers (here, with $N = \rho$ and $O = 1$). Alternatively, use the system of equations above directly (or with some variables substituted out) and stack all variables into a vector x_t to reformulate it in this form, where now F, L, G, M and H are matrices of coefficients. Indeed, if there is more than one predetermined variable like k_{t-1} in the system of equations, one will need to use such a matrix restatement of the equations anyways. More generally, z_t may also be a vector, and N and O matrices.

Anderson et al. (1996) as well as Binder and Pesaran (1997) contain detailed and general results for solving linearized systems. In most cases, the system has a solution in the form of a recursive law of motion, $x_t = Px_{t-1} + Qz_t$, for some coefficient matrices P and Q. Most models require the solution to be stable, that is, all eigenvalues of P to be less than unity in absolute value. Often, one also allows for roots equal to unity in absolute value, as this arises easily in, for example, models of international trade or with multiple agents: one may then want to think of the linear approximation as a local solution. In many models, this uniquely determines the matrix P and usually also Q.

The solutions can be found by substituting the recursive law of motion in for x_{t+1} and again for all x_t into the second-order difference equation above, exploiting $Nz_t = E_t[z_{t+1}]$ so that only x_{t-1} and z_t and some coefficient matrices remain.

Examine first the equation by matching coefficients on x_{t-1}. One obtains the equation $0 = FP^2 + GP + H$ for P. In case of a one-dimensional difference equation (as can be obtained for the example above and $x_t = k_t$), this is a quadratic equation in the feedback coefficient P, which has two solutions. The system is said to be saddle-path stable if only one of the two roots is smaller than unity in absolute value. Thus, if a stable solution is desired, this is the unique solution for P.

Generally, the equation above is a matrix quadratic equation, which can be solved per computing generalized eigenvalues or by QZ decomposition as follows. Let m be the dimensionality of x_t. Define the matrices

$$
A = \begin{bmatrix} -G & -H \\ I_m & 0_m \end{bmatrix}, B = \begin{bmatrix} F & 0_m \\ 0_m & I_m \end{bmatrix}
$$

where I_m is the m-by-m identity matrix and 0_m the m-by-m matrices of only zeros. Recall that a generalized eigenvector s with eigenvalue λ for the matrices A and B is defined as satisfying $\lambda Bs = As$. The generalized eigenvector problem reduces to the standard eigenvector problem of $B^{-1}A$, if B is invertible. If s is a generalized eigenvector with eigenvalue λ for the matrices A and B above, it can be written as $s' = [\lambda x', \underline{x'}]$ for some m-dimensional vector x. If there are m generalized eigenvalues $\lambda_1,...,\lambda_m$ together with generalized eigenvectors $s_i = [\lambda_i x_i', x_i']$ such that $C = [x_1,..., x_m]$ is of full rank, then $P = C\Lambda C^{-1}$ is a solution to the matrix quadratic equation, where

$$
\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix}
$$

is the diagonal matrix of the eigenvalues for the generalized eigenvectors used as well as of P. The system is said to be saddle-path stable if there are exactly m generalized eigenvalues smaller than

unity in absolute value. In that case, the matrix P is unique, if one requires all eigenvalues of P to be stable. If there are fewer than m eigenvalues smaller than (or equal to) unity in absolute value, then there is no solution, such that the difference equation $x_t = Px_{t-1}$ remains bounded for all x_0. In that case, the set of bounded solution is characterized by $e'x_0 = 0$ as well as $e'Qz_t = 0$ for all t for all eigenvectors e of P corresponding to explosive eigenvalues. The second of these two constraints may impose restrictions on the exogenous shock process. If there are more than m eigenvalues smaller than (or equal to) unity in absolute value, then sunspot solutions may arise, that is, there are additional solutions. In the one-dimensional case and if F is nonzero, the general solution is now given by the original equation, that is, as $x_t = -F^{-1}Gx_{t-1} - F^{-1}Hx_{t-2} - F^{-1}+(LN + M)$ $z_{t-1} + \nu t$ where ν_t is any stochastic process with $E_t[\nu_{t+1}] = 0$ and which is independent of all ε_s for $s > t$, but not necessarily independent of ε_t. Note that the recursive law of motion now includes an additional lag of the state variable, as well as the possibility for additional random influences ('sunpots') via ν_t, which are not part of the original system of equations. Farmer (1999) provides a detailed treatment of sunspots in linearized solutions.

Equivalently, consider the stacked variable $s_t' = \left[x_t', x_{t-1}'\right]$, and note that the second half of this vector is 'predetermined', that is, must be independent of all ε_s for $s > t - 1$. The linearized system can be rewritten as

$$BE_t[s_{t+1}] = As_t + \begin{bmatrix} -M - LM \\ 0 \end{bmatrix} z_t.$$

If B is invertible, the solutions can now be characterized in terms of the eigenvalues and eigenvectors of $B^{-1}A$. This is the approach taken in the classic reference of Blanchard and Kahn (1980).

Alternatively, find the QZ decomposition (or generalized Schur decomposition) of A and B (see Sims 2002), that is, find unitary matrices U and V as well as upper triangular matrices K and L such that

$$A = U'LVB = U'KV$$

(and recall that a matrix is unitary, if the product with its complex conjugate transpose is the identity matrix). Such a Schur decomposition always exists, although it may not be unique. Partition U and V into m-by-m submatrices,

$$U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}, V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}.$$

If U_{21} and V_{21} are invertible, then $P = -V_{21}^{-1}$ V_{22} solves the matrix quadratic equation. Suppose furthermore, that the QZ decomposition has been chosen so that the ratios $|L_{ii}/K_{ii}|$ are in ascending order. Furthermore, suppose $|L_{mm}/K_{mm}| < 1$. Then P is stable.

To solve for Q, given a solution to P, compare the coefficients on z_t to find $V\,\mathrm{vec}(Q) = -\mathrm{vec}(LN + M)$ where vec(.) denotes columnwise vectorization and where $V = N' \otimes F + I_k \otimes (FP + G)$ with k the dimensionality of z_t. If V is invertible, the solution is unique.

Note Many links for codes for solving dynamic stochastic models are available from QM&RBC Codes Online, Department of Economics, University of Connecticut, http://dge.repec.org/codes.html (accessed 4 September 2006). The procedure outlined above has been used in particular in the author's 'A toolkit for analyzing nonlinear economic dynamic models easily: MATLAB programs', http:// www.wiwi.hu-berlin.de/wpol/html/toolkit.htm (accessed 4 September 2006). For a discussion of the accuracy of linearized solutions, see, for example, Taylor and Uhlig (1990) and Aruoba et al. (2006).

See Also

- Multiple Equilibria in Macroeconomics
- Numerical Optimization Methods in Economics
- Prescott, Edward Christian (Born 1940)
- Real Business Cycles
- Simulation Estimators in Macroeconometrics
- Sunspot Equilibrium
- Vector Autoregressions

Bibliography

Anderson, E., E. McGrattan, L. Hansen, and T. Sargent. 1996. Mechanics of forming and estimating dynamic linear economies. In *Handbook of computational economics*, 1st ed., ed. H. Amman et al. Amsterdam: North-Holland.

Aruoba, B., J. Fernández-Villaverde, J. Rubio-Ramírez. 2006. Comparing solution methods for dynamic equilibrium economies. *Journal of Economic Dynamics and Control*.

Binder, M., and M. Pesaran. 1997. Multivariate linear rational expectations models: characterization of the nature of the solutions and their fully recursive computation. *Econometric Theory* 13: 877–888.

Blanchard, O., and C. Kahn. 1980. The solution of linear difference models under rational expectations. *Econometrica* 48: 1305–1312.

Farmer, R. 1999. *Macroeconomics of self-fulfilling prophecies*. Cambridge, MA: MIT Press.

Hansen, G. 1985. Indivisible labor and the business cycle. *Journal of Monetary Economics* 16: 309–327.

King, R., C. Plosser, and S. Rebelo. 2002. Production, growth and business cycles: technical appendix. *Computational Economics* 20: 87–116.

Kydland, F., and E. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.

Sims, C. 2002. Solving linear rational expectations models. *Computational Economics* 20: 1–20.

Taylor, J., and H. Uhlig. 1990. Solving nonlinear stochastic growth models: A comparison of alternative solution methods. *Journal of Business Economics and Statistics* 8: 1–17.

Uhlig, H. 1999. A toolkit for analysing nonlinear dynamic stochastic models easily. In *Computational methods for the study of dynamic economies*, ed. R. Marimon and A. Scott. Oxford: Oxford University Press.

Aquinas, St Thomas (1225–1274)

Barry Gordon

Keywords

Aquinas, St Thomas; Aristotle; Commutative justice; Compensation; Just price; Justice; Market price; Restitution; Usury; Value

JEL Classifications

B31

St Thomas Aquinas is generally acknowledged as the outstanding theologian of the high Middle Ages. A member of the Dominican order and a pupil of Albertus Magnus (1206–80), St Thomas taught at a number of centres including Paris, Anagni, Orvieto, Rome, Viterbo and Naples. In his research he drew on an extensive range of sources, from the Christian tradition (based on the Scriptures, the Fathers and the Roman writers) to Greek philosophy including the thought of the newly 'rediscovered' Aristotle. The writings of Aquinas are also wide-ranging, including commentaries on Aristotle's *Politics* and *Ethics.* Most celebrated among his major works is the *Summa Theologica,* which was set down between 1265 and 1273.

For St Thomas, economic reasoning is integrated with moral philosophy and the establishment of legal precepts. Analysis of economic activity is undertaken for the sake of determining appropriate standards in dealings between citizen and citizen, and so is an aspect of the inquiry into justice. The category of justice which Aquinas finds most relevant to economic life is commutative justice (from *commutatio,* that is, transaction). Hence the focal points for his economic reasoning are value and price, money and interest.

On money, St Thomas stresses its roles as a medium for the exchange of commodities and as a unit of account, that is, a standard of value or measuring rod for comparing the relative worths of exchangeable things. In his treatments of compensation for delay in repayment of a money loan and of restitution of stolen money Aquinas also recognizes that money may have economic significance when held in balance (especially when held by businessmen). The stress on money as a medium of exchange and unit of account leads to a condemnation of most forms of interest-taking as usury, hence unjust. However, the analysis of restitution and compensation help pave the way for the later acceptance by theologians of *lucrum cessans* and *damnum emergens* as phenomena offering bases for a legitimate positive rate of interest.

The just price of any commodity for St Thomas is its current market price, established in the absence of fraud or monopolistic trading

practices. It is a price established by *communiter venditur,* the price generally charged in the community concerned, rather than the price dictated by the preferences or needs of any one individual in that community. The value of a commodity will depend on subjective estimates of the utility of the good in question. It will also depend, in part, on cost of production, in that the latter influences supply conditions in any particular market. Aquinas does not achieve an effective synthesis of the utility and cost elements in his analysis of value, nor does he extend the analysis into a theory of distribution. These latter problems, however, were addressed by some of his Scholastic successors, often with reference to the analytical framework devised by St Thomas.

See Also

► Scholastic Economics

Selected Works

An English translation of Aquinas' most celebrated work is: St Thomas Aquinas, *Summa Theologiae,* translated and edited by M. Lefebure, New York: Oxford University Press, 1975. There is also a translation of one of his commentaries on Aristotle, *Commentary on the Nicomachean Ethics,* Chicago: Library of Living Catholic Thought, 1964. Selected passages from the writings of St Thomas which are of interest for economists are included in A.E. Monroe, *Early Economic Thought,* Cambridge, MA: Harvard University Press, 1924, and in A.C. Pegis, ed., *Basic Writings of St Thomas Aquinas*, 2 vols, New York: Random House, 1945. A Latin edition of Aquinas' works is: St Thomas Aquinas, *Opera Omnia,* 34 vols, ed. P. Mare and S.E. Frette, Paris: Vives, 1871–80.

Bibliography

Baldwin, J.W. 1959. *The medieval theories of the just price*. Philadelphia: American Philosophical Society.
Blaug, M., ed. 1981. *St. Thomas Aquinas*. London: Edward Elgar.
Gilchrist, J.T. 1969. *The church and economic activity in the middle ages*. London: Macmillan.
Gordon, B. 1975. *Economic analysis before Adam Smith*. London: Macmillan.
Langholm, O. 1979. *Price and value theory in the Aristotelian tradition*. Bergen: Universitetsforlaget.
Langholm, O. 1984. *The Aristotelian analysis of usury*. Bergen: Universitetsforlaget.
Langholm, O. 1998. *The legacy of scholasticism in economic thought: Antecedents of choice and power*. London: Cambridge University Press.
Noonan, J.T. 1957. *The scholastic analysis of usury*. Cambridge: Harvard University Press.
Stark, W. 1956. *The contained economy: An interpretation of medieval economic thought*. London: Blackfriars.
Torrell, J.P. 1996. *St. Thomas Aquinas. Volume I: The person and his work*. Washington, DC: Catholic University of America Press.
Torrell, J.P. 2003. *St. Thomas Aquinas. Volume II: Spiritual master*. Washington, DC: Catholic University of America Press.
Viner, J. 1978. *Religious thought and economic society*. Durham: Duke University Press.
Weisheipl, J.A. 1974. *Friar Thomas D'Aquino: His life, thought and work*. Garden City: Doubleday.
Worland, S.T. 1967. *Scholasticism and welfare economics*. Notre Dame: University of Notre Dame Press.

Arbitrage

Philip H. Dybvig and Stephen A. Ross

Abstract

The absence of arbitrage is the unifying concept for much of finance. Absence of arbitrage is more general than equilibrium because it does not require all agents to be rational. The Fundamental Theorem of Asset Pricing asserts the equivalence of absence of arbitrage, existence of a positive linear pricing rule, and existence of some hypothetical agent who prefers more to less and has an optimum. Equivalent representations of the pricing rule are the martingale measure (risk-neutral pricing), and a positive state price density. Applications of no arbitrage and these representations include Modigliani–Miller theory, option pricing, investments, and forward exchange parity.

Keywords

Arbitrage; Arrow–Debreu model; Arbitrage pricing theory; Capital asset pricing model;

Capital budgeting; Capital structure; Dividend discount model; Dividend policy; Dominance; Duality; Efficient allocation; Efficient market hypothesis; Equivalent martingale measure; Farkas' Lemma; Forward exchange, parity theory of; Fundamental theorem of asset pricing; Hahn–Banach th; Hyperplanes; Interest rates; Law of one price; Linear pricing rules; Linear programming; Markov processes; Martingales; Modigliani–Miller th; No arbitrage; Noise; Option; Option pricing; Pricing rule representation th; Purchasing power parity; Risk premium; Risk-neutral probabilities; Separation theorems; State price density; State spaces; Trade costs; von Neumann and Morgenstern

JEL Classifications
G0

An arbitrage opportunity is an investment strategy that guarantees a positive payoff in some contingency with no possibility of a negative payoff and with no net investment. By assumption, it is possible to run the arbitrage possibility at arbitrary scale; in other words, an arbitrage opportunity represents a money pump. A simple example of arbitrage is the opportunity to borrow and lend costlessly at two different fixed rates of interest. Such a disparity between the two rates cannot persist: arbitrageurs will drive the rate together.

The modern study of arbitrage is the study of the implications of assuming that no arbitrage opportunities are available. Assuming no arbitrage is compelling because the presence of arbitrage is inconsistent with equilibrium when preferences increase with quantity. More fundamentally, the presence of arbitrage is inconsistent with the existence of an optimal portfolio strategy for any competitive agent who prefers more to less, because there is no limit to the scale at which an individual would want to hold the arbitrage position. Therefore, in principle, absence of arbitrage follows from individual rationality of a single agent. One appeal of results based on the absence of arbitrage is the intuition that absence of arbitrage is more primitive than equilibrium, since only relatively few rational agents are needed to bid away arbitrage opportunities, even in the presence of a sea of agents driven by 'animal spirits'.

The absence of arbitrage is very similar to the zero economic profit condition for a firm with constant returns to scale (and no fixed factors). If such a firm had an activity which yielded positive profits, there would be no limit to the scale at which the firm would want to run the activity, and no optimum would exist. The theoretical distinction between a zero profit condition and the absence of arbitrage is the distinction between commerce, which requires production, and trading under the price system, which does not. In practice, the distinction blurs. For example, if gold is sold at different prices in two markets, there is an arbitrage opportunity but it requires production (transportation of the gold) to take advantage of the opportunity. Furthermore, there are almost always costs to trading in markets (for example, brokerage fees), and therefore a form of costly production is required to convert cash into a security. For the purposes of this article, we will tend to ignore production. In practical applications the necessity of production will weaken the implications of absence of arbitrage and may drive a wedge between what the pure absence of arbitrage would predict and what actually occurs.

The assertion that two perfect substitutes (for example, two shares of stock in the same company) must trade at the same price is an implication of no arbitrage that goes under the name of the law of one price. While the law of one price is an immediate consequence of the absence of arbitrage, it is not equivalent to the absence of arbitrage. An early use of a no-arbitrage condition employed the law of one price to help explain the pattern of prices in the foreign exchange and commodities markets.

Many economic arguments use the absence of arbitrage implicitly. In discussions of purchasing power parity in international trade, for example, presumably it is an arbitrage possibility that forces the spot exchange rate between currencies to equal the relative prices of common baskets of (traded) goods. Similarly, the statement that the possibility of repackaging implies linear prices in competitive product markets is essentially a no-arbitrage argument.

Early Uses of the Law of One Price

The parity theory of forward exchange based on the law of one price was first formulated by Keynes (1923) and developed further by Einzig (1937). Let s denote the current spot price of, say, euros, in terms of dollars, and let f denote the forward price of euros one year in the future. The forward price is the price at which agreements can be struck currently for the future delivery of euros with no money changing hands today. Also, let r_s and r_m denote the one year dollar and euro interest rates, respectively. To prevent an arbitrage possibility from developing, these four prices must stand in a particular relation.

To see this, consider the choices facing a holder of dollars. The holder can lend the dollars in the domestic market and realize a return of r_s one year from now. Alternatively, the investor can purchase euros on the spot market, lend for one year in the German market, and convert the euros back into dollars one year from now at the fixed forward rate. By undertaking the conversion back into dollars in the forward market, the investor locks in the prevailing forward rate, f. The results of this latter path are a return of

$$f(1 + r_m)/s$$

dollars one year from now. If this exceeds $1 + r_s$, then the foreign route offers a sure higher return than domestic lending. By borrowing dollars at the domestic rate r_s and lending them in the foreign market, a sure profit at the rate

$$f(1 + r_m)/s - (1 + r_s)$$

can be made with no net investment of funds. Alternatively, if

$$f(1 + r_m)/s - (1 + r_s) < 0,$$

the arbitrage works in reverse. By borrowing in euros, investing in dollars, and buying euros forward, a sure profit at the rate

$$(1 + r_s) - f(1 + r_m)/s$$

can be made with no investment in funds.

Thus, the prevention of arbitrage will enforce the forward parity result,

$$(1 + r_s)/(1 + r_m) = f/s.$$

This result takes on many different forms as we look across different markets. In a commodity market with costless storage, for example, an arbitrage opportunity will arise if the following relation does not hold:

$$f \leq s(1 + r).$$

In this equation, f is the currently quoted forward rate for the purchase of the commodity – for example, silver, one year from now – s is the current spot price, and r is the interest rate. More generally, if c is the up-front proportional carrying cost, including such items as storage costs, spoilage and insurance, absence of arbitrage ensures that

$$f \leq s(1 + c)(1 + r).$$

(We normally would expect these relations to hold with equality in a market in which positive stocks are held at all points in time, and perhaps with inequality in a market which may not have positive stocks just before a harvest. However, proving equality is based on equilibrium arguments, not on the absence of arbitrage, since to short the physical commodity one must first own a positive amount.)

The above applications of the absence of arbitrage (via the law of one price) share the common characteristic of the absence of risk. The law of one price is less restrictive than the absence of arbitrage because it deals only with the case in which two assets are identical but have different prices. It does not cover cases in which one asset dominates another but may do so by different amounts in different states. The most interesting applications of the absence of arbitrage are to be found in uncertain situations, where this distinction may be important.

The Fundamental Theorem of Asset Pricing

The absence of arbitrage is implied by the existence of an optimum for any agent who prefers

more to less. The most important implication of the absence of arbitrage is the existence of a positive linear pricing rule, which in many spaces including finite state spaces is the same as the existence of positive state prices that correctly price all assets. Taken together with their converses, we refer collectively to these results as the *Fundamental Theorem of Asset Pricing*. Traditionally, the emphasis has been on the linear pricing rule as an implication of the absence of arbitrage. Including the existence of an optimum (introduced in the version of this article in the first edition of *The New Palgrave*) is useful both because it reminds us why we are interested in arbitrage, and because the converse tells us that, absent other restrictions, consistency with equilibrium is equivalent to the absence of arbitrage. We state the theorem verbally here; the formal meanings of the words and the proof are given later in this section.

Theorem (*Fundamental Theorem of Asset Pricing*) The following are equivalent:

(i) absence of arbitrage;
(ii) existence of a positive linear pricing rule;
(iii) existence of an optimal demand for some agent who prefers more to less.

Beja (1971) was one of the first to emphasize explicitly the linearity of the asset pricing function, but he did not link it to the absence of arbitrage. Beja simply assumed that equilibrium prices existed and observed 'that equilibrium properties require that the functional q be linear' where q is a functional that assigns a price or value to a risky cash flow. The first statement and proof that the absence of arbitrage implied the existence of non-negative state space prices and, more generally, of a positive linear operator that could be used to value risky assets appeared in Ross (1976a, 1978). Besides providing a formal analysis, Ross showed that there was a pricing rule that prices *all* assets and not just those actually marketed. (In other words, the linear pricing rule could be extended from the marketed assets to all hypothetical assets defined over the same set of states.) The advantage of this extension is that the domain of the pricing function does not depend on the set of marketed assets. We will largely follow Ross's analysis with some modern improvements.

Linearity for pricing means that the price functional or operator q satisfies the ordinary linear condition of algebra. If we let x and y be two random payoffs and we let q be the operator that assigns values to prospects, then we require that

$$q(ax + by) \quad = \quad aq(x) + bq(y),$$

where a and b are arbitrary constants. Of course, for many spaces (including a finite state space), any linear functional can be represented as a sum or integral across states of state prices times quantities.

To simplify proofs in this article, we will make the assumption that there are finitely many states, each of which occurs with positive probability, and that all claims purchased today pay off at a single future date. Let Θ denote the state space,

$$\Theta = \{1, \ldots, m\},$$

where there are m states and the state of nature θ occurs with probability π_θ. Applying q to the 'indicator' asset e_θ whose payoff is 1 in state θ and 0 otherwise, we can define a price q_θ for each state θ as the value of e_θ;

$$q_\theta \quad = \quad q(e_\theta).$$

Now, if there were linearity, the value of any payoff, x, could be written as

$$q(x) \quad = \quad \sum_\theta q_\theta x_\theta.$$

Of course, this argument presupposes that $q(e_\theta)$ is well defined, which is a strong assumption if e_θ is not marketed.

We want to make a statement about the conditions under which all marketed assets can be priced by such a linear pricing rule q. We assume that there is a set of n marketed assets with a corresponding price vector, p. Asset i has a terminal payoff X_{θ_i} (inclusive of dividends, and so on) in state of nature θ. The matrix $X \equiv [X_{\theta_i}]$ denotes the state space tableau whose columns correspond to assets and whose rows correspond to states.

Lower-case x represents the random vector of terminal payoffs to the various securities. An arbitrage opportunity is a portfolio (vector) η with two properties. It does not cost anything today or in a state in the future. And, it has a positive payoff either today or in some state in the future (or both). We can express the first property as a pair of vector inequalities. The initial cost is not greater than zero, which is to say that it uses no wealth and may actually generate some,

$$p\eta \leq 0, \tag{1}$$

and its random payoff later is never negative,

$$X\eta \geq 0. \tag{2}$$

(We use the notation that $\geq$ denotes greater or equal in each component, $>$ denotes $\geq$ and greater in some component, and $\gg$ denotes greater in all components. Note that writing the price of $X\eta$ as $p\eta$ for arbitrary η embodies an assumption that investment in marketed assets is divisible.) The second property says that the arbitrage portfolio η has a strict inequality, either in (1) or in some component of (2). We can express both properties together as

$$X_*\eta \;=\; \begin{bmatrix} -p \\ X \end{bmatrix}\eta > 0. \tag{3}$$

Here, we have stacked the net payoff today on top of the vector of payoffs at the future date. This is in the spirit of the Arrow–Debreu model in which consumption in different states, commodities, points of time and so forth, are all considered components of one large consumption vector.

The absence of arbitrage is simply the condition that no η satisfies (3). A consistent positive linear pricing rule is a vector of state prices $q \gg 0$ that correctly prices all marketed assets, that is, such that

$$p \;=\; qX. \tag{4}$$

We have now collected enough definitions to prove the first half (that (i) $\Leftrightarrow$ (ii)) of the Fundamental Theorem of Asset Pricing.

Theorem (*first half of the Fundamental Theorem of Asset Pricing*) There is no arbitrage if and only if there exists a consistent positive linear pricing rule.

Proof The proof that having a consistent positive linear pricing rule precludes arbitrage is simple, since any arbitrage opportunity gives a direct violation of (4). Let η be an arbitrage opportunity. By (4),

$$p\eta \;=\; qX\eta,$$

or equivalently

$$0 \;=\; -p\eta + q(X\eta) \;=\; [1q]X_*\eta.$$

By definition of an arbitrage opportunity (3) and positivity of q, we have a contradiction.

The proof that the absence of arbitrage implies the existence of a consistent positive linear pricing rule is more subtle and requires a separation theorem. The mathematical problem is equivalent to Farkas' Lemma of the alternative and to the basic duality theorem of linear programming. We will adopt an approach that is analogous to the proof of the second theorem of welfare economics that asserts the existence of a price vector which supports any efficient allocation, by separating the aggregate Pareto optimal allocation from all aggregate allocations corresponding to Pareto preferable allocations. Here we will find a price vector that 'supports' an arbitrage-free allocation by separating the net trades from the set of free lunches (the positive orthant).

The absence of arbitrage is equivalent to the requirement that the linear space of net trades defined by

$$s \;\equiv\; \{y|\ \text{for some}\ \eta, y \;=\; X_*\eta\}, \tag{5}$$

does not intersect the positive orthant $\mathscr{R}^m_+ + 1 = \{y|y \geq 0\}$ except at the origin, that is $S \cap \mathscr{R}^m_+ + 1 = \{0\}$.

Since S is a subspace (and is therefore a convex closed cone), a simple separation theorem (Karlin 1959, Theorem B3.5) implies that there exists a nonzero vector q_* such that for all $y \in S$ and all $z \in \mathscr{R}^m_+ + 1$, $z \neq 0$, we must have

$$q_* z > 0 \geq q_* y. \tag{6}$$

Letting z be each of the unit vectors in turn, the first inequality in (6) implies that q_* is a strictly positive vector.

Since S is a subspace, the second inequality in (6) must hold with equality for all $y \in S$ Define

$$q \equiv (q_{*2}, q_{*3}, \ldots, q_{*n})/q_{*1}.$$

Since $q_* \gg 0$, likewise $q \gg 0$.

Dividing the second equality in (6) (which we now know to be an equality) by q_{*1} and expanding using the definition of X_* [from (3)], we have that

$$0 = -p + qX,$$

or

$$p = qX,$$

which shows that q is a consistent positive linear pricing rule.

Before we can prove the second half of the pricing theorem, we need to define the maximization problem faced by a typical investor. In this problem, all we really need to assume is that more is preferred (strictly) to less, that is, that increasing initial consumption or random consumption later in one or more states always leads to a preferred outcome. In fact, this is literally all we need: we do not need completeness or even transitivity of preferences, let alone a utility function representation or any restriction to a functional form. However, for concreteness, we will write down preferences using a state-dependent utility function of consumption now and in the future. The assumption that the investor prefers more to less is satisfied if the utility function in each state is increasing in consumption at both dates.

The state-dependent restriction implies that the maximization problem faced by a particular agent is the maximization of the expectation of the state-dependent utility function $u_0(\cdot, \cdot)$ of initial wealth and terminal wealth, given initial wealth w_0 and the possibility of trading in the security market. Then the maximization problem faced by a typical agent is the unconstrained choice of a vector α of portfolio weights to maximize

$$\sum_0 \pi_0 u_0\left[w_0 - p\alpha, (X\alpha)_0\right].$$

The quantity $p\alpha$ is the price of the portfolio, and therefore $w_0 - p\alpha$ is the residual amount of the initial wealth available for initial consumption. The preferences of the agent are said to be increasing if each $u_0(\cdot, \cdot)$ is (strictly) increasing in both arguments. Saying the agent prefers more to less is just another way of saying that preferences are increasing.

Here is the rest of the proof of the Fundamental Theorem of Asset Pricing.

Theorem (*second half of the Fundamental Theorem of Asset Pricing*) There is no arbitrage if and only if there exists some (at least hypothetical) agent with increasing preferences whose choice problem has a maximum.

Proof If there is an arbitrage opportunity, η, then clearly the choice problem for an agent with increasing preferences cannot have a maximum, since for every α,

$$\sum_\theta \pi_\theta u_\theta\left\{w_0 - p(\alpha + k\eta), [X(\alpha + k\eta)]_\theta\right\}$$

increases as k increases.

Conversely, if there is no arbitrage, by the first half of the Fundamental Theorem of Asset Pricing (proven earlier), there exist a consistent positive linear pricing rule q. Let $w_0 = 0$ and $\alpha = 0$. Consider the particular utility function

$$\begin{aligned} u_{*\theta}(c_0, c_1) \equiv & -\exp[-(c_0 - w_0)] \\ & -(q_\theta/\pi_\theta)\exp(-c_1). \end{aligned} \tag{7}$$

Each function $u_{*\theta}$ is strictly increasing and also happens to be strictly concave, infinitely differentiable, and additively separable over time. Using $p = qX$, it is easy to show that this utility function satisfies the first-order conditions for a maximum, which are necessary and sufficient by concavity. (Note: by a more complicated argument, it can be shown that the von Neumann–Morgenstern 'state independent' utility function $-\exp(-c_0) - \exp(-c_1)$ has a maximum, but the maximum will not necessarily be achieved at $\alpha = 0$).

As should be clear from the proof, it is not really important what class of preference we use, so long as all agents having preferences in the class prefer more to less and the class includes the particular preferences used in the proof (which are additive over states and time, increasing, concave, and infinitely differentiable).

Recent research on arbitrage, starting with Ross (1978) and Harrison and Kreps (1979), has focused on extending these results to more general state spaces in which there are many time periods and, more importantly, infinitely many states. In these spaces, deriving a positive linear pricing rule for marketed claims is still straightforward (one can prove the algebraic linearity condition and positivity directly from the no-arbitrage condition), but extending the pricing rule from the priced claims to all non-marketed claims requires some sort of extension theorem, such as a Hahn–Banach theorem. Obtaining a truly general result is complicated by the fact that the positive orthant is not typically an open set in these general spaces, and openness is a condition of the Hahn–Banach theorems. One part of the result that goes through in general is the implication that existence of an optimum implies existence of a linear pricing rule: so long as preferences are continuous in our topology, the preferred set will be open, and the linear pricing rule will be a hyperplane that separates the optimum from the preferred set.

Alternative Representations of Linear Pricing Rules

There are many equivalent ways of representing a linear pricing rule. Which representation is simplest depends on the context. In one representation, the price is the expected value under artificial 'risk-neutral' probabilities discounted at the riskless rate. (The risk-neutral probability measure is also referred to as an equivalent martingale measure.) In another representation, the price is the expectation of the quantity times the state price density, which is the state price per unit probability. In yet another representation, the price is the expected value discounted at a risk- adjusted rate. The purpose of this section is to show the fundamental equivalence of these representations.

The motive for using a particular representation is usually found in the study of intertemporal models or models with a continuum of states. Nonetheless, we will continue our formal analysis of the single-period model with finitely many states, leaving the more general discussion of the merits of the various approaches until afterwards. Now, we have already seen the basic linear pricing rule representation. For any portfolio α,

$$p\alpha = qX\alpha = \sum_{\theta} q_{\theta}(X\alpha)_{\theta}, \tag{8}$$

that is, the sum across states of state price times the payoff.

The risk-neutral or martingale representation asserts the existence of a vector Π of artificial probabilities and a shadow riskless rate r such that

$$\begin{aligned} p\alpha &= (1+r)^{-1}\Pi X\alpha \\ &= (1+r)^{-1}E_n(x\alpha), \end{aligned} \tag{9}$$

that is, the expectation E_{Π} of the payoff under the risk-neutral (martingale) probabilities Π, discounted at the riskless rate. It is easy to see the shadow riskless rate is equal to the riskless rate if one exists. The risk neutral approach is trivially equivalent to the positive linear pricing rule approach. Simply let

$$\Pi = q/\sum_{\theta} q_{\theta} \tag{10}$$

and

$$(1+r)^{-1} = \sum_{\theta} q_{\theta} \tag{11}$$

For the converse, let

$$q = (1+r)^{-1} = \sum_{\theta} q_{\theta} \tag{12}$$

Therefore, the existence of a positive linear pricing rule is the same as the existence of positive risk-neutral probabilities. (The risk-neutral measure is

equivalent to the original probability measure, that is, Π has the same null sets as α. Here, that is simply the requirement that the list of states with positive probability is the same for both measures.)

A third approach emphasizes the role of the state price density, ρ_0. In this case, the price is given by

$$p\alpha = \sum_{\theta} \pi_\theta \rho_\theta (X\alpha)_\theta = E(\rho x \alpha). \quad (13)$$

To see that this is equivalent to the linear pricing rule, simply let

$$\rho_\theta = q_\theta / \pi_\theta, \quad (14)$$

or, conversely, let

$$q_\theta = \rho_\theta \pi_\theta. \quad (15)$$

Clearly, p is positive in all states if and only if q is.

We have shown the equivalence of these three approaches. This equivalence is stated in the following theorem.

Theorem (*Pricing Rule Representation Theorem*) The following are equivalent:

- existence of a positive linear pricing rule;
- existence of positive risk-neutral probabilities and an associated riskless rate (the martingale property);
- existence of a positive state price density.

The remaining representation is that the value is equal to the terminal value discounted at a risk-adjusted interest rate r_a.

$$p\alpha = (1 + r_a)^{-1} E(x\alpha) \quad (16)$$

While this might at first appear to be inconsistent with the other representations, the risk-adjusted rate r_a is typically proportional to the covariance of return ($=x\alpha/p\alpha$) with some random variable, and consequently solving this equation for px yields a linear rule. (See Beja 1971; Rubinstein 1976, for general results concerning pricing rules using covariances.) For example, in the capital asset pricing model,

$$r_a = r + \lambda \mathrm{cov}(x\alpha/p\alpha, r_m), \quad (17)$$

where r_m is the random return on the market and λ is the market price of risk. Solving these two equations for px, we obtain

$$p\alpha = (1 + r)^{-1} E[x\alpha\{1 - \lambda[r_m - E(\mathrm{r}_m)]\}], \quad (18)$$

which is certainly linear in $x\alpha$. The subtle question is whether or not this is positive, and this hinges on whether the market return can get larger than $E(r_m) + 1/\lambda$ (Dybvig and Ingersoll 1982). In any case, the important observation is that the basic form of the representation is linear even if verification of positivity depends on the exact form of the risk premium.

Now we return to the question of the comparative advantages of the various representations. The risk-neutral or martingale representation was first employed by Cox and Ross (1976a) for use in option pricing problems and was later developed more formally by Harrison and Kreps (1979) and a number of others. The risk- neutral representation is particularly useful for problems of valuation or optimization without reference to individual preferences, since under the martingale probabilities we can ignore risk altogether and maximize discounted expected value. In fact, for some problems this approach tells us that risk-neutral results generalize immediately to worlds where risk is priced. However, this approach tends to be complicated when preferences are introduced, since von Neumann–Morgenstern (state independent) preferences under ordinary probabilities become state dependent under the martingale probabilities. As an aside, we note that, in intertemporal contexts in which the interest rate is stochastic, the price is the risk-neutral expectation of the future value discounted by the rolled-over spot rate (which is stochastic).

The state price density representation (Cox and Leland 2000; Dybvig 1980, 1988) is most useful when we want to look at choice problems. Samuelson (1947) emphasized the value of deriving equilibrium conditions from first- and second-order conditions for optimization. In asset pricing

the first-order condition for an agent with von Neumann–Morgenstern preferences is that the agent's marginal utility of consumption is proportional to a consistent state price density (not necessarily unique) for the security market (Dybvig and Ross 1982). (Note that if there is a nonatomic continuum of states, the state price density will typically be well-defined even though all primitive states have probability zero and state price zero.) For the CAPM, this fact was used implicitly by Sharpe (1964) and Lintner (1965), and was made explicit by Dybvig and Ingersoll (1982).

The representation of discounting expected returns using a risk-adjusted rate is most useful when we can get some independent assessment of the risk premium involved. Otherwise, it is needlessly complicated, since the price appears not only on the left-hand side of the equation but also in the denominator on the right-hand side. Discounting using a risk-adjusted rate is usually the method of choice for capital budgeting, since the risk adjustment is usually determined from comparables (for example, from past returns on assets in similar firms). For capital budgeting, there may also be a pedagogical advantage that (so far) it has been easier to communicate to practitioners than the other methods. Furthermore, focusing on the risk-adjusted discount rate sharpens the comparison of competing approaches (such as the capital asset pricing model and the dividend discount model).

It is useful to note how the various representations evolve over time. State prices are simply the product of state prices over sub-periods. For example, for $t < s < T$, the state price of a state at T given the state at t is equal to the state price of the state at T given the state at s times the state price of the state at s given the state at t. (The state at s is determined by the state at T given the pervasive assumption of perfect recall, that is, the assumption that the family of sigma-algebras is increasing. If we use some reduced specification of the state – as when looking at Markov processes – the state price is the product of the two, summed over all possible intermediate states.)

The martingale representation yields a price equal to the expected value under the martingale measure of the product of the terminal value times a discount factor that corresponds to rolling over shortest maturity default-free bonds. This representation makes particularly clear the interaction between term structure effects and other effects. If there is a significant term structure, the discount factor is random, and we cannot ignore the interplay between term structure risk and random terminal value unless the terminal value of the asset under consideration is independent of interest rates (under the martingale measure). If the terminal value is independent of interest rate movements, then the value of the asset today is the risk- neutral expected terminal value of the asset discounted at the riskless discount factor (which equals the risk-neutral expected discount factor from rolling over shorts).

The state price density has an evolution over time similar to that of the state price, namely, the state price density over a long interval is the product of the state price density over short intervals. Since the state price density equals the state price divided by the probability, the ratio of the two evolutions gives us a relation involving only probabilities, which is Bayes' law.

Finally, the discounted expected value approach is more complicated than the others. The exact evolution over time depends on whether uncertainty is multiplicative, linear, a distributed lag, or whatever. This difficulty is usually overlooked in capital budgeting applications, which is probably not so bad in practice, given the imprecision of our estimates of risk premia and future cash flows.

Modern Results Based on the Absence of Arbitrage

Most of modern finance is based on either the intuitive or the actual theory of the absence of arbitrage. In fact, it is possible to view absence of arbitrage as the one concept that unifies all of finance (Ross 1978). In this section, we will try to provide a sample of how arbitrage arguments are used in diverse areas in finance. We will touch on applications in option pricing, corporate finance, asset pricing and efficient markets.

The efficient market hypothesis says that the price of an asset should fully reflect all available information. The intuition behind this hypothesis is that, if the price does not fully reflect available information, then there is a profit opportunity available from buying the asset if the asset is underpriced or from selling it if it is overpriced. Clearly this is consistent with the intuition of the absence of arbitrage, even if what we have here is only an approximate arbitrage possibility, that is, a large profit at little risk. Approximate arbitrage is always profitable to a risk-neutral investor. More generally, the issue is clouded somewhat by qsts of risk tolerance and what is the appropriate risk premium. Happily, empirical violation of efficiency of the market (for example, in event studies) is not significantly affected by the procedure for measuring the risk premium (Brown and Warner 1980, 1985). Therefore, an empirical violation of efficiency is an approximate arbitrage opportunity that presumably would be attractive at large scale to many investors.

The Modigliani–Miller propositions tell us that, in perfect capital markets, changing capital structure or dividend policy without changing investment is a matter of irrelevance to the shareholders. The original proofs of the Modigliani–Miller propositions used the law of one price and assumed the presence of a perfect substitute for the firm that was altering its capital structure. As an illustration of the Fundamental Theorem of Asset Pricing, Ross (1978) demonstrated that these propositions could be derived directly from the existence of a positive linear pricing rule.

To illustrate this argument, consider the proposition that the total value of the firm does not depend on the capital structure. The original argument assumed that there is another identical firm. If we change the financing of our firm, then the value of holding a portfolio of all the parts will give a final payoff equal to that of the identical firm, and must therefore have the same value under the law of one price. Alternatively, suppose that there exists a positive linear pricing rule q. Let x represent the total terminal value of a firm in a one-period model and xi the payoff to financial claim i on the assets of the firm. Then the sum of all the payoffs must add up to the total terminal value.

$$x = \sum_i x_i \tag{19}$$

Using the positive linear operator, q, which values assets, we have that the value of the firm,

$$\begin{aligned} v &\equiv \sum\nolimits_i q(x_i) = q\left(\sum\nolimits_i x_i\right) \\ &= q(x_i), \end{aligned} \tag{20}$$

which is independent of the number of structure of the financial claims.

Note that both proofs make an implicit assumption that goes beyond what absence of arbitrage promises, namely, that changing the capital structure of the firm does not change the way in which prices are formed in the economy. In the original proof this is the assumption that the other firm's price will not change when the firm changes its capital structure. In the linear pricing rule proof this is the assumption that the state price vector q does not change.

Another application of the absence of arbitrage is to asset pricing. The most obvious application is the derivation of the arbitrage pricing theory (Ross 1976a, b). We will consider the special case without asset-specific noise. Assume that the mechanism generating the per dollar investment rates of return for a set of assets is given by

$$R_i = E_i + \beta_{i1} f_1 + \ldots + \beta_{ik} f_k, \quad i = 1, \ldots, n. \tag{21}$$

where E_i is the expected rate of return on asset i per dollar invested and f_i is an exogenous factor. This form is an exact factor generating mechanism (as opposed to an approximate one with an additional asset specific mean zero term).

Applying the pricing operator, q, to Eq. (21) we have that

$$\begin{aligned} 1 &= q(1 + R_i) \\ &= q(1 + E_i + \beta_{i1} f_1 + \ldots + \beta_{ik} f_k) \\ &= q(1 + E_i) + \beta_{i1}\, q(f_1) + \ldots + \beta_{ik}\, q(f_k) \\ &= (1 + E_i)/(1 + r) + \beta_{ik} q(f_1) + \ldots + \beta_{ik}\, q(f_k), \end{aligned}$$

which implies that

$$E_i - r = \lambda_1 \beta_{i1} + \ldots + \lambda_k \beta_{ik}, \tag{22}$$

where $\lambda_j \equiv -(1 + r)q(f_i)$ is the risk premium associated with factor j. Equation (22) is the basic equation of the arbitrage pricing theory. We have derived it using absence of exact arbitrage in the absence of asset-specific noise. More general derivations account for asset-specific noise and use absence of approximate arbitrage.

The most important paper in option pricing, Black and Scholes (1973), is based on the absence of arbitrage, as is the whole literature it has generated. At any point in time, the option is priced by duplicating the value one period later using a portfolio of other assets, and assigning a value using the law of one price. We will illustrate this procedure using the binomial process studies by Cox et al. (1979). During each period, the stock price either goes up by 20 per cent or it goes down by 10 per cent, and for simplicity we take the riskless rate to be zero. Assume that we are one period from the maturity of a call option with an exercise price of \$100, and that the stock price is now \$100 (the call is at the money).

How much is the option worth? To figure this out, we must find a portfolio of the stock and the bond that gives the same terminal value. This is the solution of two linear equations (one for each state) in two unknowns (the two portfolio weights). Explicitly, the terminal call value is the larger of 0 and the stock price less 100. In the good state, the stock value will be \$120 and the option will be worth \$20. In the bad state, the stock price will be \$90 and the option will be worthless. If α_S is the amount of stock and α_B the amount of \$100 face bond to hold in the duplicating portfolio, then we have that

$$20 = 120\alpha_S + 100\alpha_B$$

to duplicate the option value in the good state, and

$$0 = 90\alpha_S + 100\alpha_B$$

to duplicate the option value in the bad state. The solution to the two equations is given by

$$\alpha_S = 2/3\alpha_B = -3/5.$$

Therefore, each option is equivalent to holding 2/3 shares of stock and shorting (borrowing) 3/5 bonds. By the law of one price, the option value is the value of this portfolio, or $100\alpha_S + 100\alpha_B = 6\ 2/3$. In this context, we used arbitrage to value the option exactly. More generally, if less is known about the form of the stock price process, absence of arbitrage still places useful restrictions on the option price (Merton 1973; Cox and Ross 1976b). For example, the price of a call option is less than the current stock price, and the price of a European put option is no smaller than the present value of the stock price less the current stock price.

Absence of arbitrage also implies a surprising feature of the behaviour of long interest rates in the limit as maturity increases. Let $V(t, T)$ denote the zero-coupon bond price, namely, the price at t of a riskless claim for \$1 at T. Equivalently, we can describe bond prices in terms of the zero-coupon rate $z(t, T)$ where $V(t, T) = 1/(1 + z(t, T))T - t$. Defining the long zero-coupon rate, $zL(t) \equiv \lim T \uparrow \infty z(t, T)$; absence of arbitrage implies that the probability is zero that this rate will ever fall. This is because the bond price today is an average of bond prices tomorrow weighted by (positive) state prices, and the bond price in any state declines asymptotically at the rate $zL(t)$ in that state tomorrow. Thus, the weighted average of prices today declines at a rate equal to the smallest rate under our maintained assumption of finitely many states (and perhaps more slowly given infinitely many states). As a consequence, $zL(t)$ at time t is always less than or equal to its value $zL(t)$ at any future date, $s > t$, in every realization (with probability one). For details see Dybvig et al. (1996).

Dominance is a useful concept to combine with the absence of arbitrage. A dominance argument gives features of a strategy that are optimal independent of preferences and, often, independent of distributions as well. For example, when we write the payoff on a call as $\max(S - X, 0)$, we are implicitly assuming it is a chosen strategy to exercise the option when it is in the money and not to exercise it when it is out of the money. Absent frictions, this is a dominant strategy and the assumption is without loss of generality. A more subtle dominance argument, relying on the absence of frictions and on a non-negative riskless

rate, gives the classical result that an American call option (which can be exercised at or before maturity) has the same value as the corresponding European call option (which can only be exercised at maturity), because waiting to exercise is a dominant strategy (Merton 1973; Cox and Ross 1976b). Another dominance argument can be used to show that it is optimal to exercise certain reload options used in executive compensation again and again, whenever they are in the money (Dybvig and Loewenstein 2003).

An alternative to option pricing by arbitrage is to use a 'preference-based' model and price options using the first-order conditions of an agent (Rubinstein 1976). While using this alternative approach is very convenient in some contexts, the Fundamental Theorem of Asset Pricing tells us that we are not really doing anything different, and that the two approaches are simply two different ways of making the same assumption. The same point is true of the distinction some authors have made between the 'equilibrium' derivations of the arbitrage pricing theory and the 'arbitrage' derivations: there is no substance in this distinction. One derivation may give a tighter approximation than another, but all derivations require similar assumptions in one form or another.

See Also

- ▶ Finance
- ▶ Modigliani–Miller Theorem
- ▶ Options
- ▶ Present Value

Bibliography

Beja, A. 1971. The structure of the cost of capital under uncertainty. *Review of Economic Studies* 38: 359–368.

Black, F., and M.S. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.

Brown, S., and J. Warner. 1980. Measuring security price performance. *Journal of Financial Economies* 8: 205–258.

Brown, S., and J. Warner. 1985. Using daily stock returns: The case of event studies. *Journal of Financial Economics* 14: 3–31.

Cox, J., and H. Leland. 2000. On dynamic investment strategies. *Journal of Economic Dynamics and Control* 24: 1859–1880.

Cox, J., and S.A. Ross. 1976a. The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3: 145–166.

Cox, J., and S.A. Ross. 1976b. A survey of some new results in financial option pricing theory. *Journal of Finance* 31: 383–402.

Cox, J., S. Ross, and M. Rubinstein. 1979. Option pricing: A simplified approach. *Journal of Financial Economics* 7: 229–263.

Dybvig, P. 1980. Some new tools for testing market efficiency and measuring mutual fund performance. Unpublished manuscript.

Dybvig, P. 1988. Distributional analysis of portfolio choice. *Journal of Business* 61: 369–393.

Dybvig, P., and J. Ingersoll Jr. 1982. Mean-variance theory in complete markets. *Journal of Business* 55: 233–251.

Dybvig, P., and M. Loewenstein. 2003. Employee reload options: Pricing, hedging, and optimal exercise. *Review of Financial Studies* 16: 145–171.

Dybvig, P., and S. Ross. 1982. Portfolio efficient sets. *Econometrica* 50: 1525–1546.

Dybvig, P., J. Ingersoll, and S.A. Ross. 1996. Long forward and zero-coupon rates can never fall. *Journal of Business* 69: 1–25.

Einzig, P. 1937. *The theory of forward exchange*. London: Macmillan.

Harrison, J.M., and D. Kreps. 1979. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20: 381–408.

Karlin, S. 1959. *Mathematical methods and theory in games, programming, and economics*. Reading: Addison-Wesley.

Keynes, J.M. 1923. *A tract on monetary reform*. London: Macmillan.

Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47: 13–37.

Merton, R. 1973. Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4: 141–183.

Ross, S.A. 1976a. Return, risk and arbitrage. In *Risk and return in finance*, ed. I. Friend and J. Bicksler. Cambridge, MA: Ballinger.

Ross, S.A. 1976b. The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13: 341–360.

Ross, S.A. 1978. A simple approach to the valuation of risky streams. *Journal of Business* 51: 453–475.

Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics and Management Science* 7: 407–425.

Samuelson, P. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Sharpe, W. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.

Arbitrage Pricing Theory

Gur Huberman and Zhenyu Wang

Abstract

Focusing on asset returns governed by a factor structure, the APT is a one-period model, in which preclusion of arbitrage over static portfolios of these assets leads to a linear relation between the expected return and its covariance with the factors. The APT, however, does not preclude arbitrage over dynamic portfolios. Consequently, applying the model to evaluate managed portfolios contradicts the no-arbitrage spirit of the model. An empirical test of the APT entails a procedure to identify features of the underlying factor structure rather than merely a collection of mean-variance efficient factor portfolios that satisfies the linear relation.

Keywords

Arbitrage; Arbitrage pricing theory; Arrow–Debreu security pricing; Asset allocation; Asset pricing; Black–Scholes model; Capital asset pricing model; Cost of capital; Factor models; Generalized method of moments; Hilbert space techniques; Mean- variance efficiency; Portfolio analysis; Stochastic discount factor

JEL Classifications

G12

The arbitrage pricing theory (APT) was developed primarily by Ross (1976a, b). It is a one-period model in which every investor believes that the stochastic properties of returns of capital assets are consistent with a factor structure. Ross argues that, if equilibrium prices offer no arbitrage opportunities over static portfolios of the assets, then the expected returns on the assets are approximately linearly related to the factor loadings. (The factor loadings, or betas, are proportional to the returns' covariances with the factors.) The result is stated in section "A Formal Statement".

Ross's (1976a) heuristic argument for the theory is based on the preclusion of arbitrage. This intuition is sketched in section "Intuition". Ross's formal proof shows that the linear pricing relation is a necessary condition for equilibrium in a market where agents maximize certain types of utility. The subsequent work, which is surveyed below, derives either from the assumption of the preclusion of arbitrage or the equilibrium of utility maximization. A linear relation between the expected returns and the betas is tantamount to an identification of the stochastic discount factor (SDF). Sections "No-Arbitrage Models" and "Utility-Based Arguments", respectively, review this literature.

The APT is a substitute for the capital asset pricing model (CAPM) in that both assert a linear relation between assets' expected returns and their covariance with other random variables. (In the CAPM, the covariance is with the market portfolio's return.) The covariance is interpreted as a measure of risk that investors cannot avoid by diversification. The slope coefficient in the linear relation between the expected returns and the covariance is interpreted as a risk premium. Such a relation is closely tied to mean-variance efficiency, which is reviewed in section "Mean-Variance Efficiency".

Section "Mean-Variance Efficiency" also points out that an empirical test of the APT entails a procedure to identify at least some features of the underlying factor structure. Merely stating that some collection of portfolios (or even a single portfolio) is mean-variance efficient relative to the mean-variance frontier spanned by the existing assets does not constitute a test of the APT, because one can always find a mean-variance efficient portfolio. Consequently, as a test of the APT it is not sufficient to merely show that a set of factor portfolios satisfies the linear relation between the expected return and its covariance with the factors portfolios.

A sketch of the empirical approaches to the APT is offered in section "Empirical Tests", while section "Specification of Factors" describes various procedures to identify the underlying factors. The large number of factors proposed in the literature and the variety of statistical or ad hoc

procedures to find them indicate that a definitive insight on the topic is still missing.

Finally, section "Applications" surveys the applications of the APT, the most prominent being the evaluation of the performance of money managers who actively change their portfolios. Unfortunately, the APT does not necessarily preclude arbitrage opportunities over dynamic portfolios of the existing assets. Therefore, the applications of the APT in the evaluation of managed portfolios contradict at least the spirit of the APT, which obtains price restrictions by assuming the absence of arbitrage.

A Formal Statement

The APT assumes that investors believe that the $n \times 1$ vector, r, of the single-period random returns on capital assets satisfies the factor model

$$r = \mu + \beta f + e, \tag{1}$$

where e is an $n \times 1$ vector of random variables, f is a k $\times$ 1 vector of random variables (factors), μ is an $n \times 1$ vector and β is an $n \times k$ matrix. With no loss of generality, normalize (1) to make $E[f] = 0$ and $E[e] = 0$,where $E[\cdot]$ denotes expectation and 0 denotes the matrix of zeros with the required dimension. The factor model (1) implies $E[r] = \mu$.

The mathematical proof of the APT requires restrictions on β and the covariance matrix $\Omega = E[ee']$. An additional customary assumption is that $E[e/f] = 0$, but this assumption is not necessary in some of the APT's developments.

The number of assets, n, is assumed to be much larger than the number of factors, k. In some models, n is infinity or approaches infinity. In this case, representation (1) applies to a sequence of capital markets; the first n assets in the $(n + 1)$st market are the same as the assets in the nth market and the first n rows of the matrix β in the $(n + 1)$st market constitute the matrix β in the nth market.

The APT asserts the existence of a constant a such that, for each n, the inequality

$$(\mu - X\lambda)Z^{-1}(\mu - X\lambda) \leq a \tag{2}$$

holds for a $(k + 1) \times 1$ vector λ, and an $n \times n$ positive definite matrix Z. Here, $X = (l, \beta)$, in which l is an $n \times 1$ vector of ones. Let λ_0 be the first component of λ and λ_1 consists of the rest of the components. If some portfolio of the assets is risk- free, then λ_0 is the return on the risk-free portfolio. The positive definite matrix Z is often the covariance matrix $E[ee']$. *Exact* arbitrage pricing obtains if (2) is replaced by

$$\mu = X\lambda = {}_l\lambda_0 + \beta\lambda_1. \tag{3}$$

The vector λ_1 is referred to as the risk premium, and the matrix β is referred to as the beta or loading on factor risk.

The interpretation of (2) is that each component of μ depends approximately linearly on the corresponding row of β. This linear relation is the same across assets. The approximation is better, the smaller the constant a; if $a = 0$, the linear relation is exact and (3) obtains.

Intuition

The intuition behind the model draws from the intuition behind Arrow–Debreu security pricing. A set of k fundamental securities spans all possible future states of nature in an Arrow–Debreu model. Each asset's payoff can be described as the payoff on a portfolio of the fundamental k assets. In other words, an asset's payoff is a weighted average of the fundamental assets' payoffs. If market clearing prices allow no arbitrage opportunities, then the current price of each asset must equal the weighted average of the current prices of the fundamental assets.

The Arrow–Debreu intuition can be couched in terms of returns and expected returns rather than payoffs and prices. If the unexpected part of each asset's return is a linear combination of the unexpected parts of the returns on the k fundamental securities, then the expected return of each asset is the same linear combination of the expected returns on the k fundamental assets.

To see how the Arrow–Debreu intuition leads from the factor structure (1) to exact arbitrage

pricing (3), set the idiosyncratic term e on the right-hand side of (1) equal to zero. Translate the k factors on the right-hand side of (1) into the k fundamental securities in the Arrow–Debreu model. Then (3) follows immediately.

The presence of the idiosyncratic term e in the factor structure (1) makes the model more general and realistic. It also makes the relation between (1) and (3) more tenuous. Indeed, 'no arbitrage' arguments typically prove the weaker (2). Moreover, they require a weaker definition of arbitrage (and therefore a stronger definition of no arbitrage) in order to get from (1) to (2).

The proofs of (2) augment the Arrow–Debreu intuition with a version of the law of large numbers. That law is used to argue that the average effect of the idiosyncratic terms is negligible. In this argument, the independence among the components of e is used. Indeed, the more one assumes about the (absence of) contemporaneous correlations among the component of e, the tighter the bound on the deviation from exact APT.

No-Arbitrage Models

Huberman (1982) formalizes Ross's (1976a) heuristic argument. A portfolio v is an $n \times 1$ vector. The cost of the portfolio v is $v'l$, the income from it is $v'r$, and its return is $v'r/v'l$ (if its cost is not zero). Huberman defines arbitrage as the existence of zero- cost portfolios such that a subsequence $\{w\}$ satisfies

$$\lim_{n\to\infty} E[w'r] = \infty \text{ and } \lim_{n\to\infty} \text{var}[w'r] = 0, \quad (4)$$

where var[·] denotes variance. The first requirement in (4) is that the expected income associated with w becomes large as the number of assets increases. The second requirement in (4) is that the risk (as measured by the income's variance) vanishes as the number of assets increases. Accordingly, a sequence of capital markets offers no arbitrage if there is no subsequence $\{w\}$ of zero-cost portfolios that satisfy (4).

Huberman shows that, if the factor model (1) holds and if the covariance matrix $E[ee']$ is diagonal for all n and uniformly bounded, then the absence of arbitrage implies (2) with $Z = I$ and a finite bound a. The idea of his proof is as follows. Consider the orthogonal projection of the vector μ on the linear space spanned by the columns of X:

$$\mu = X\hat{\lambda} + \alpha, \quad (5)$$

where $\alpha'X = 0$ and $\hat{\lambda}$ is a $k \times 1$ vector. The projection implies

$$\alpha'\alpha = \min_{\lambda} (\mu - X\lambda)'(\mu - X\lambda). \quad (6)$$

A violation of (2) is the existence of a subsequence of $\{\alpha'\alpha\}$ that approaches infinity. The vector α is often referred to as a pricing error and it can be used to construct arbitrage. For any scalar h, the portfolio $w = h\alpha$ has zero cost because the first column of X is ι. The factor model (1) and the projection (5) imply $E[w'r] = h(\alpha'\alpha)$ and var $[w'r] = h^2(\alpha'E[ee']\alpha)$. If σ^2 is the upper bound of the diagonal elements of $E[ee']$, then var $[w'r] \le h^2(\alpha'\alpha)\sigma^2$. If h is chosen to be $(\alpha'\alpha)^{-2/3}$, then $E[w'r] = (\alpha'\alpha)^{1/3}$ and $\text{var}[w'r] \le (\alpha'\alpha)^{-1/3}\sigma^2$, which imply that (4) is satisfied by a subsequence of the zero-cost portfolios $\{(\alpha'\alpha)^{-2/3}\alpha\}$.

Using the no-arbitrage argument, the exact APT can be proven to hold in the limit for well-diversified portfolios. A portfolio w is well diversified if $w'_l = 1$ and $\text{var}[w'e] = 0$, that is, if the portfolio's return contains only factor variance. A sequence of portfolios, $\{w\}$, is well diversified if $w'_l = 1$ and $\lim_{n\to\infty}\text{var}[w'e] = 0$. Suppose there are m sequences of well-diversified portfolios and m is a fixed number larger than $k + 1$. For each n, let W bean $n \times m$ matrix, in which each column is one of the well-diversified portfolios. The exact APT holds in the limit for the well-diversified portfolios if and only if there exists a sequence of $k \times 1$ vectors, $\{\lambda\}$, such that

$$\lim_{n\to\infty} \left(W'\mu - \tilde{X}\lambda\right)' \left(W'\mu - \tilde{X}\lambda\right) = 0, \quad (7)$$

where $\tilde{X} = (\mathrm{j}, W'\beta)$ and j is an $m \times 1$ vector of ones. The projection of $W'\mu$ on the columns of $\tilde{X}$

gives $W'\mu = \tilde{X}\,\hat{\lambda} + \alpha$, in which $\alpha'\tilde{X} = 0$. If Eq.(7) does not hold, a subsequence of α satisfies $\alpha'\alpha > \delta$ for some positive constant δ. This sequence of α can be used to construct arbitrage as follows. For any scalar h, define a portfolio as $v = hW\alpha$, which is then costless because $v_l' = h\alpha' W_l' = h\alpha' j = 0$. It follows from $\alpha'\tilde{X} = 0$ that $E[v'\ r] = h\alpha'\alpha$ and var $[v'\ r] = h^2\alpha' W'\ E[ee']W\alpha$. If h is chosen to be $(\alpha'\ W'\ E[ee']W\alpha)^{1/3}$, then var$[v'\ r] = h^{-1}$. Since $\{w\}$ is well-diversified and $E[ee']$ is diagonal and uniformly bounded, it follows that $\lim_{n\to\infty} h = \infty$. This implies that portfolio sequence $\{v\}$ is arbitrage because it satisfies (4).

Ingersoll (1984) generalizes Huberman's result, showing that the factor model, uniform boundedness of the elements of β and no arbitrage imply (2) with $Z = E[ee']$, which is not necessarily diagonal. A variant of Ingersoll's argument is as follows. Write the positive definite matrix Z as the product $Z = UU'$, where U is an $n\ n \times n$ on-singular matrix. Then, consider the orthogonal projection of the vector $U{-}1\mu$ on the column space of $U^{-1}X$:

$$U^{-1}\mu = U^{-1}X\lambda + \alpha, \tag{8}$$

where $\alpha' U^{-1}X = 0$. The rest of the argument is similar to those presented earlier.

Chamberlain and Rothschild (1983) employ Hilbert space techniques to study capital markets with (possibly infinitely) many assets. The preclusion of arbitrage implies the continuity of the cost functional in the Hilbert space. Let L equal the maximum eigenvalue of the limit covariance matrix $E[ee']$ and d equal the supremum of all the ratios of expectation to standard deviation of the incomes on all costless portfolios with a nonzero weight on at least one asset. Chamberlain and Rothschild demonstrate that (2) holds with $a = Ld^2$ and $Z = I$ if asset prices allow no arbitrage.

With two additional assumptions, Chamberlain (1983) provides explicit lower and upper bounds on the left-hand side of (2). He further shows that exact arbitrage pricing obtains if and only if there is a well-diversified portfolio on the mean- variance frontier. The first of his additional assumptions is that all the factors can be represented as limits of traded assets. The second additional assumption is that the variances of incomes on any sequence of portfolios that are well diversified in the limit and that are uncorrelated with the factors converge to zero.

Utility-Based Arguments

In utility-based arguments, investors are assumed to solve the following problem:

$$\max_{c_0, c_T, w} E[u(c_0, c_T)] \text{ subject to } c_0 \le b - w_l' \text{ and } c_T \le w_r', \tag{9}$$

where b is the initial wealth, and $u(c_0, c_T)$ is a utility function of initial and terminal consumption c_0 and c_T. The utility function is assumed to increase with initial and with terminal consumption. The first order condition is

$$E[rM] = l, \tag{10}$$

where $M = (\partial u/\partial c_T)/(\partial u/\partial c_0)$.The random variable M satisfying (10) is referred to as the stochastic discount factor (SDF) by Hansen and Jagannathan (1991, 1997). Substitution of the factor model (1) into the first order condition gives

$$\mu = {}_l\lambda^l 0 + \beta\lambda_1 + \alpha, \tag{11}$$

where $\lambda_0 = 1/E[M]$, $\lambda_1 = -E[fM]/E[M]$ and $\alpha = -E[eM]/E[M]$. It follows from (11) that

$$(\mu - X\lambda)'(\mu - X\lambda) = \alpha'\alpha, \tag{12}$$

where $X = (l,\ \beta)$ and $\lambda = \left(\lambda_0, \lambda_1'\right)'$.

Clearly, the APT (2) holds for $Z = I$and a if $\alpha'\alpha$ is uniformly bounded by a. Ross (1976a) is the first to set up an economy in which $\alpha'\alpha$ is uniformly bounded. The exact APT (3) holds if and only if

$$E[eM] = 0. \tag{13}$$

If the SDF is a linear function of the factors, then Eq. (13) holds. Conversely, if Eq. (13) holds,

there exists an SDF, which is a linear function of factors, such that Eq. (10) is satisfied. However, the SDF does not have to be a linear function of factors for the purpose of obtaining the exact APT. A nonlinear function, $M = g(f)$, of factors for the SDF would also imply (13) under the assumption $E[e/f] = 0$.

Connor (1984) shows that, if the market portfolio is well diversified, then every investor holds a well-diversified portfolio (that is, a $k + 1$ fund separation obtains; the funds are associated with the factors and with the risk-free asset, which Connor assumes to exist). With this, the first order condition of any investor implies exact arbitrage pricing in a competitive equilibrium.

Connor and Korajczyk (1986) extend Connor's previous work to a model with investors who have better information about returns than most other investors. The former class of investors is sufficiently small, so the pricing result remains intact and it is used to derive a test of the superiority of information of the allegedly better informed investors.

Connor and Korajczyk (1988) extend Connor's single-period model to a multi- period model. They assume that the capital assets are the same in all periods, that each period's cash payoffs from these assets obey a factor structure, and that competitive equilibrium prices are set as if the economy had a representative investor who maximizes exponential utility. They show that exact arbitrage pricing obtains with time-varying risk premium (but, similar to Stambaugh 1983, with constant factor loadings.)

Chen and Ingersoll (1983) argue that, if a well-diversified portfolio exists and it is the optimal portfolio of some utility-maximizing investor, then the first order condition of that investor implies exact arbitrage pricing.

Dybvig (1983) and Grinblatt and Titman (1983) consider the case of finite assets and provide explicit bounds on the deviations from exact arbitrage pricing. These bounds are functions of the per capita asset supplies, individual bounds on absolute risk aversion, variance of the idiosyncratic risk, and the interest rate. To derive his bound, Dybvig assumes that the support of the distribution of the idiosyncratic term e is bounded below, that each investor's coefficient of absolute risk aversion is non-increasing and that the competitive equilibrium allocation is unconstrained Pareto optimal. To derive their bound, Grinblatt and Titman require a bound on a quantity related to investors' coefficients of absolute risk aversion and the existence of k independent, costless and well diversified portfolios.

Mean-Variance Efficiency

The APT was developed as a generalization of the CAPM, which asserts that the expectations of assets' returns are linearly related to their covariances (or betas, which in turn are proportional to the covariances) with the market portfolio's return. Equivalently, the CAPM says that the market portfolio is mean-variance efficient in the investment universe containing all possible assets. If the factors in (1) can be identified with traded assets, then exact arbitrage pricing (3) says that a portfolio of these factors is mean-variance efficient in the investment universe consisting of the assets r.

Huberman and Kandel (1985b), Jobson and Korkie (1982, 1985) and Jobson (1982) note the relation between the APT and mean-variance efficiency. They propose likelihood-ratio tests of the joint hypothesis that a given set of random variables are factors in model (1) and that exact arbitrage pricing (3) obtains. Kan and Zhou (2001) point out a crucial typographical error in Huberman and Kandel (1985b). Peñaranda and Sentana (2004) study the close relation between the Huberman and Kandel's spanning approach and the celebrated volatility bounds in Hansen and Jagannathan (1991).

Even when the factors are not traded assets, (3) is a statement about mean- variance efficiency: Grinblatt and Titman (1987) assume that the factor structure (1) holds and that a risk-free asset is available. They identify k traded assets such that a portfolio of them is mean-variance efficient if and only if (3) holds. Huberman et al. (1987) extend the work of Grinblatt and Titman by characterizing the sets of k traded assets with that property and show that these assets can be described as

portfolios if and only if the global minimum variance portfolio has non-zero systematic risk. To find these sets of assets, one must know the matrices $\beta\beta'$ and $E[ee']$.If the latter matrix is diagonal, factor analysis produces an estimate of it, as well as an estimate of $\beta\beta'$.

The interpretation of (3) as a statement about mean-variance efficiency contributes to the debate about the testability of the APT. (Shanken 1982, 1985, and Dybvig and Ross 1985, however, discuss the APT's testability without mentioning that (3) is a statement about mean-variance efficiency.) The theory's silence about the factors' identities renders any test of the APT a joint test of the pricing relation and the correctness of the factors. As a mean-variance efficient portfolio always exists, one can always find 'factors' with respect to which (3) holds. In fact, any single portfolio on the frontier can serve as a 'factor'.

Thus, finding portfolios which are mean-variance efficient – or failure to find them – neither supports nor contradicts the APT. It is the factor structure (1) which, combined with (3), provides refutable hypotheses about assets' returns. The factor structure (1) imposes restrictions which, combined with (3), provide refutable hypotheses about assets' returns. The factor structure suggests looking for factors with two properties: (*a*) their time-series movements explain a substantial fraction of the time-series movements of the returns on the priced assets, and (*b*) the unexplained parts of the time series movements of the returns on the priced assets are approximately uncorrelated across the priced assets.

Empirical Tests

Empirical work inspired by the APT typically ignores (2) and instead studies exact arbitrage pricing (3). This type of work usually consists of two steps: an estimation of factors (or at least of the matrix β) and then a check to see whether exact arbitrage pricing holds. In the first step, researchers typically use the following regression model to estimate the parameters in the factor model:

$$r_t = \alpha + \beta f_t + e_t, \tag{14}$$

where *rt*, *ft* and *et* are the realization of the variables in period *t*. The factors observed in empirical studies often have a non-zero mean, denoted by δ. Let T be the total number of periods and Σ the summation over $t = 1, \ldots, T$. The ordinary least-square (OLS) estimates are

$$\hat{\mu} = \frac{1}{T}\Sigma r_t \text{ and } \hat{\delta} = \frac{1}{T}\Sigma f_t \tag{15}$$

$$\hat{\beta} = \left(\Sigma(r_t - \hat{\mu})\ (f_t - \hat{\mu})'\right) \times \left(\Sigma\left(f_t - \hat{\delta}\right)\left(f_t - \hat{\delta}\right)'\right)^{-1} \tag{16}$$

$$\hat{\alpha} = \hat{\mu} - \hat{\beta}\,\hat{\delta} \tag{17}$$

$$\hat{\Omega} = \frac{1}{T}\Sigma \hat{e}_t \hat{e}_t' \text{ where } \hat{e}_t = r_t - \hat{\alpha} - \hat{\beta} f_t. \tag{18}$$

These are also maximum-likelihood estimators if the returns and factors are independent across time and have a multivariate normal distribution.

In the second step, researchers may use the exact pricing (3) and (14) to obtain the following restricted version of the regression model,

$$r_t = \iota\lambda_0 + \beta(f_t + \lambda_1) + e_t. \tag{19}$$

Under the assumption that returns and factors follow identical and independent normal distributions, the maximum-likelihood estimators are

$$\overline{\beta} = \left(\Sigma\left(r_t - \iota\overline{\lambda}_0\right)\left(f_t + \overline{\lambda}_1\right)'\right) \times \left(\Sigma\left(f_t + \overline{\lambda}_1\right)\left(f_t + \overline{\lambda}_1\right)'\right)^{-1} \tag{20}$$

$$\overline{\Omega} = \frac{1}{T}\Sigma \overline{e}_t \overline{e}_t' \text{ where } \overline{e}_t = r_t - \iota\overline{\lambda}_0 - \overline{\beta}\left(f_t + \overline{\lambda}_1\right) \tag{21}$$

$$\overline{\lambda} = \left(\overline{X}'\overline{\Omega}^{-1}\overline{X}\right)^{-1}\overline{X}'\overline{\Omega}^{-1}\left(\hat{\mu} - \overline{\beta}\hat{\delta}\right) \text{ where } \overline{X} = \left(\iota, \overline{\beta}\right). \tag{22}$$

These estimators need to be solved simultaneously from the above three equations. Notice that $\overline{\beta}$ and $\overline{\Omega}$ are the OLS estimators in (19) for a given $\overline{\lambda}$. The last equation shows that $\overline{\lambda}$ is the

generalized least-square estimator in the cross-sectional regression of $\hat{\mu} - \overline{\beta}\hat{\delta}$ on $\overline{X}$ with $\overline{\Omega}$ being the weighting matrix. To test the restriction imposed by the exact APT, researchers use the likelihood-ratio statistic,

$$\mathrm{LR} = T\left(\log\left|\overline{\Omega}\right| - \log\left|\hat{\Omega}\right|\right), \tag{23}$$

which follows a χ^2 distribution with $n - k - 1$ degrees of freedom when the number of observations, T, is very large. When factors are payoffs of traded assets or a risk-free asset exists, the exact APT imposes more restrictions. For these cases, Campbell et al. (1997, ch. 6) provide an overview. If the observations of returns and factors do not follow independent normal distribution, similar tests can be carried out using the generalized method of moments (GMM). Jagannathan and Wang (2002) and Jagannathan et al. (2002) provide an overview of the application of the GMM for testing asset pricing models including the APT.

Interest is sometimes focused only on whether a set of specified factors are priced or on whether their loadings help explain the cross section of expected asset returns. For this purpose, most researchers study the cross-sectional regression model

$$\hat{\mu} = \hat{X}\lambda + v \text{ or } \hat{\mu} = \iota\lambda_0 + \hat{\beta}\lambda_1 + v, \tag{24}$$

where $\hat{X} = \left(\iota, \hat{\beta}\right)$ and v is an $n \times 1$ vector of errors for this equation. The OLS estimator of λ in this regression is tested to see whether it is different from zero. To test this specification, asset characteristics z, such as firm size, that are correlated with mean asset returns are added to the regression:

$$\hat{\mu} = \iota\lambda_0 + \hat{\beta}\lambda_1 + z\lambda_2 + v. \tag{25}$$

A significant λ_1 and insignificant λ_2 are viewed as evidence in support of the specified factors being part of the exact APT. Black et al. (1972) and Fama and MacBeth (1973) pioneered this cross-sectional approach to test the CAPM. Chen et al. (1986) used it to test the exact APT. Shanken (1992) and Jagannathan and Wang (1998) developed the statistical foundations of the cross- sectional tests. The cross-sectional approach is now a popular tool for analysing risk premiums on the loadings of proposed factors.

Specification of Factors

The tests outlined above are joint tests that the matrix β is correctly estimated and that exact arbitrage pricing holds. Estimation of the factor loading matrix β entails at least an implicit identification of the factors. The three approaches listed below have been used to identify factors.

The first consists of an algorithmic analysis of the estimated covariance matrix of asset returns. For instance, Roll and Ross (1980), Chen (1983) and Lehman and Modest (1988) use factor analysis, and Chamberlain and Rothschild (1983) and Connor and Korajczyk (1986, 1988) recommend using principal component analysis.

The second approach is one in which a researcher starts at the estimated covariance matrix of asset returns and uses his judgement to choose factors and subsequently estimate the matrix β. Huberman and Kandel (1985a) note that the correlations of stock returns of firms of different sizes increase with a similarity in size. Therefore, they choose an index of small firms, one of medium-size firms and one of large firms to serve as factors. In a similar vein, Fama and French (1993) use the spread between the stock returns of small and large firms as one of their factors. Echoing the findings of Rosenberg et al. (1984), Chan et al. (1991) and Fama and French (1992) observe that expected stock returns and their correlations are also related to the ratio of book-to-market equity. Based on these observations, Fama and French (1993) add the spread between stock returns of value and growth firms as another factor.

The third approach is purely judgemental in that it is one in which the researcher primarily uses his intuition to pick factors and then estimates the factor loadings and checks whether they explain the cross-sectional variations in estimated expected returns (that is, he checks (3)).

Chan et al. (1985) and Chen et al. (1986) select financial and macroeconomic variables to serve as factors. They include the following variables: the return on an equity index, the spread of short- and long-term interest rates, a measure of the private sector's default premium, the inflation rate, the growth rates of industrial production and the aggregate consumption. Based on economic intuition, researchers continue to add new factors, which are too many to enumerate here.

The first two approaches are implemented to conform to the factor structure underlying the APT: the first approach by the algorithmic design and the second because researchers check that the factors they use indeed leave the unexplained parts of asset returns almost uncorrelated. The third approach is implemented without regard to the factor structure. Its attempt to relate the assets' expected returns to the covariance of the assets' returns with other variables is more in the spirit of Merton's (1973) inter-temporal CAPM than in the spirit of the APT.

The empirical work cited above examines the extent to which the exact APT (with whatever factors are chosen) explains the cross-sectional variation in assets' mean returns better than the CAPM. It also examines the extent to which other variables – usually those that include various firm characteristics – have marginal explanatory power beyond the factor loadings to explain the cross section of assets' mean returns. The results usually suggest that the APT is a useful model in comparison with the CAPM. (Otherwise, they would probably have gone unpublished.) However, the results are mixed when the alternative is firm characteristics. Researchers who introduce factors tend to report results supporting the APT with their factors and test portfolios. Nevertheless, different tests and construction of portfolios often reject the proposed APT. For example, Fama and French (1993) demonstrate that exact APT using their factors holds for portfolios constructed by sorting stocks on firm size and book-to-market ratio, whereas Daniel and Titman (1997) demonstrate that the same APT does not hold for portfolios that are constructed by sorting stocks further on the estimated loadings with respect to Fama and French's factors.

The APT often seems to describe the data better than competing models. It is wise to recall, however, that the purported empirical success of the APT may well be due to the weakness of the tests employed. Some questions come to our mind: which factors capture the data best; what is the economic interpretation of the factors; what are the relations among the factors that different researchers have reported? As any test of the APT is a joint test that the factors are correctly identified and that the linear pricing relation holds, a host of competing theories exist side by side under the APT's umbrella. Each fails to reject the APT but has its own factor identification procedure. The number of factors, as well as the methods of factor construction, is exploding. The multiplicity of competing factor models indicates ignorance of the true factor structure of asset returns and suggests a rich and challenging research agenda.

Applications

The APT lends itself to various practical applications due to its simplicity and flexibility. The three areas of applications critically reviewed here are: asset allocation, the computation of the cost of capital, and the performance evaluation of managed funds.

The application of the APT in asset allocation is motivated by the link between the factor structure (1) and mean-variance efficiency. Since the structure with k factors implies the existence of k assets that span the efficient frontier, an investor can construct a mean-variance efficient portfolio with only k assets. The task is especially straightforward when the k factors are the payoffs of traded securities. When k is a small number, the model reduces the dimension of the optimization problem. The use of the APT in the construction of an optimal portfolio is equivalent to imposing the restriction of the APT in the estimation of the mean and covariance matrix involved in the mean-variance analysis. Such a restriction increases the reliability of the estimates because it reduces the number of unknown parameters.

If the factor structure specified in the APT is incorrect, however, the optimal portfolio

constructed from the APT will not be mean-variance efficient. This uncertainty calls for adjusting, rather than restricting, the estimates of mean and covariance matrix by the APT. The degree of this adjustment should depend on investors' prior belief in the model. Pastor and Stambaugh (2000) introduce the Bayesian approach to achieve this adjustment. Wang (2005) further shows that the Bayesian estimation of the return distribution results in a weighted average of the distribution restricted by the APT and the unrestricted distribution matched to the historical data.

The proliferation of APT-based models challenges an investor engaging in asset allocation. In fact, Wang (2005) argues that investors averse to model uncertainty may choose an asset allocation that is not mean-variance efficient for any probability distributions estimated from the prior beliefs in the model.

Being an asset pricing model, the APT should lend itself to the calculation of the cost of capital. Elton et al. (1994) and Bower and Schink (1994) used the APT to derive the cost of capital for electric utilities for the New York State Utility Commission. Elton, Gruber and Mei specify the factors as unanticipated changes in the term structure of interest rates, the level of interest rates, the inflation rate, the GDP growth rate, changes in foreign exchange rates, and a composite measure they devise to measure changes in other macro factors. In the meantime, Bower and Schink use the factors suggested by Fama and French (1993) to calculate the cost of capital for the Utility Commission. However, the Commission did not adopt any of the above-mentioned multi-factor models but used the CAPM instead (see DiValentino 1994).

Other attempts to apply the APT to compute the cost of capital include Bower et al. (1984), Goldenberg and Robin (1991) who use the APT to study the cost of capital for utility stocks, and Antoniou et al. (1998) who use the APT to calculate the cost of equity capital when examining the impact of the European exchange rate mechanism. Different studies use different factors and consequently obtain different results, a reflection of the main drawback of the APT – the theory does not specify what factors to use. According to Green et al. (2003), this drawback is one of the main reasons that the US Federal Reserve Board has decided not to use the APT to formulate the imputed cost of equity capital for priced services at Federal Reserve Banks.

The application of asset pricing models to the evaluation of money managers was pioneered by Jensen (1968). When using the APT to evaluate money managers, the managed funds' returns are regressed on the factors, and the intercepts are compared with the returns on benchmark securities such as Treasury bills. Examples of this application of the APT include Busse (1999), Carhart (1997), Chan et al. (2002), Caiet al. (1997), Elton et al. (1996), Mitchell and Pulvino (2001), and Pastor and Stambaugh (2002).

The APT is a one-period model that delivers arbitrage-free pricing of existing assets (and portfolios of these assets), given the factor structure of their returns. Applying it to price derivatives on existing assets or to price trading strategies is problematic, because its stochastic discount factor is a random variable which may be negative. Negativity of the SDF in an environment which permits derivatives leads to a pricing contradiction, or arbitrage. Consider, for instance, the price of an option that pays its holder whenever the SDF is negative. Being a limited liability security, such an option should have a positive price, but applying the SDF to its payoff pattern delivers a negative price. (The observation that the stochastic discount factor of the CAPM may be negative is in to Dybvig and Ingersoll 1982, who also studied some of the implications of this observation.)

Trading and derivatives on existing assets are closely related. Famously, Black and Scholes (1973) show that dynamic trading of existing securities can replicate the payoffs of options on these existing securities. Therefore, one should be careful in interpreting APT-based excess returns of actively managed funds because such funds trade rather than hold on to the same portfolios. Examples of interpretations of asset management techniques as derivative securities include Merton (1981) who argues that market-timing strategy is an option, Fung and Hsieh (2001) who show that hedge funds using trend-following strategies behave like a look-back

straddle, and Mitchell and Pulvino (2001) who demonstrate that merger arbitrage funds behave like an uncovered put.

Motivated by the challenge of evaluating dynamic trading strategies, Glosten and Jagannathan (1994) suggest replacing the linear factor models with the Black–Scholes model. Wang and Zhang (2005) study the problem extensively and develop an econometric methodology to identify the problem in factor-based asset pricing models. They show that the APT with many factors is likely to have large pricing errors over actively managed funds, because empirically these models deliver SDFs which allow for arbitrage over derivative-like payoffs.

It is ironic that some of the applications of the APT require extensions of the basic model which violate its basic tenet – that assets are priced as if markets offer no arbitrage opportunities.

See Also

- ▶ Arbitrage
- ▶ Capital Asset Pricing Model
- ▶ Factor Models

Bibliography

Antoniou, A., I. Garrett, and R. Priestley. 1998. Calculating the equity cost of capital using the APT: The impact of the ERM. *Journal of International Money and Finance* 14: 949–965.

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.

Black, F., M. Jensen, and M. Scholes. 1972. The capital-asset pricing model: Some empirical tests. In *Studies in the theory of capital markets*, ed. M. Jensen. New York: Praeger Publishers.

Bower, R., and G. Schink. 1994. Application of the Fama–French model to utility stocks. *Financial Markets, Institutions and Instruments* 3: 74–96.

Bower, D., R. Bower, and D. Logue. 1984. Arbitrage pricing and utility stock returns. *Journal of Finance* 39: 1041–1054.

Busse, J. 1999. Volatility timing in mutual funds: Evidence from daily returns. *Review of Financial Studies* 12: 1009–1041.

Cai, J., K. Chan, and T. Yamada. 1997. The performance of Japanese mutual funds. *Review of Financial Studies* 10: 237–273.

Campbell, J., A. Lo, and C. MacKinley. 1997. *The econometrics of financial markets*. Princeton: Princeton University Press.

Carhart, M. 1997. On persistence in mutual fund performance. *Journal of Finance* 52: 57–82.

Chamberlain, G. 1983. Funds, factors and diversification in arbitrage pricing models. *Econometrica* 51: 1305–1323.

Chamberlain, G., and M. Rothschild. 1983. Arbitrage, factor structure, and mean variance analysis on large asset markets. *Econometrica* 51: 1281–1304.

Chan, K., N. Chen, and D. Hsieh. 1985. An exploratory investigation of the firm size effect. *Journal of Financial Economics* 14: 451–471.

Chan, L., Y. Hamao, and J. Lakonishok. 1991. Fundamentals and stock returns in Japan. *Journal of Finance* 46: 1739–1764.

Chan, L., H. Chen, and J. Lakonishok. 2002. On mutual fund investment styles. *Review of Financial Studies* 15: 1407–1437.

Chen, N. 1983. Some empirical tests of the theory of arbitrage pricing. *Journal of Finance* 38: 1393–1414.

Chen, N., and J. Ingersoll. 1983. Exact pricing in linear factor models with infinitely many assets: A note. *Journal of Finance* 38: 985–988.

Chen, N., R. Roll, and S. Ross. 1986. Economic forces and the stock markets. *Journal of Business* 59: 383–403.

Connor, G. 1984. A unified beta pricing theory. *Journal of Economic Theory* 34: 13–31.

Connor, G., and R. Korajczyk. 1986. Performance measurement with the arbitrage pricing theory: A framework for analysis. *Journal of Financial Economics* 15: 373–394.

Connor, G., and R. Korajczyk. 1988. Risk and return in an equilibrium APT: Application of a new test methodology. *Journal of Financial Economics* 21: 213–254.

Daniel, K., and S. Titman. 1997. Evidence on the characteristics of cross sectional variation in stock returns. *Journal of Finance* 52: 1–33.

DiValentino, L. 1994. Preface. *Financial Markets, Institutions and Instruments* 3: 6–8.

Dybvig, P. 1983. An explicit bound on deviations from APT pricing in a finite economy. *Journal of Financial Economics* 12: 483–496.

Dybvig, P., and J. Ingersoll. 1982. Mean-variance theory in complete markets. *Journal of Business* 55: 233–251.

Dybvig, P., and S. Ross. 1985. Yes, the APT is testable. *Journal of Finance* 40: 1173–1188.

Elton, E., M. Gruber, and J. Mei. 1994. Cost of capital using arbitrage pricing theory: A case study of nine New York utilities. *Financial Markets, Institutions and Instruments* 3: 46–73.

Elton, E., M. Gruber, and C. Blake. 1996. Survivorship bias and mutual fund performance. *Review of Financial Studies* 9: 1097–1120.

Fama, E., and K. French. 1992. The cross-section of expected stock returns. *Journal of Finance* 47: 427–486.

Fama, E., and K. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33: 3–56.

Fama, E., and J. MacBeth. 1973. Risk, return and equilibrium: Empirical tests. *Journal of Political Economy* 71: 607–636.

Fung, W., and D. Hsieh. 2001. The risks in hedge fund strategies: Theory and evidence from trend followers. *Review of Financial Studies* 14: 313–341.

Glosten, L., and R. Jagannathan. 1994. A contingent claim approach to performance evaluation. *Journal of Empirical Finance* 1: 133–160.

Goldenberg, G., and A. Robin. 1991. The arbitrage pricing theory and cost-of-capital estimation: The case of electric utilities. *Journal of Financial Research* 14: 181–196.

Green, E., J. Lopez, and Z. Wang. 2003. Formulating the imputed cost of equity for priced services at Federal Reserve Banks. *Economic Policy Review* 9: 55–58.

Grinblatt, M., and S. Titman. 1983. Factor pricing in a finite economy. *Journal of Financial Economics* 12: 495–507.

Grinblatt, M., and S. Titman. 1987. The relation between mean-variance efficiency and arbitrage pricing. *Journal of Business* 60: 97–112.

Hansen, L., and R. Jagannathan. 1991. Implications of security market data for models of dynamic economies. *Journal of Political Economy* 99: 225–262.

Hansen, L., and R. Jagannathan. 1997. Assessing specification errors in stochastic discount factor models. *Journal of Finance* 52: 557–590.

Huberman, G. 1982. A simple approach to arbitrage pricing. *Journal of Economic Theory* 28: 183–191.

Huberman, G., and S. Kandel. 1985a. *A size based stock returns model*. Center for Research in Security Prices Working Paper 148, University of Chicago.

Huberman, G., and S. Kandel. 1985b. *Likelihood ratio tests of asset pricing and mutual fund separation*. Center for Research in Security Prices Working Paper 149, University of Chicago.

Huberman, G., S. Kandel, and R. Stambaugh. 1987. Mimicking portfolios and exact arbitrage pricing. *Journal of Finance* 42: 1–9.

Ingersoll, J. 1984. Some results in the theory of arbitrage pricing. *Journal of Finance* 39: 1021–1039.

Jagannathan, R., and Z. Wang. 1998. An asymptotic theory for estimating beta-pricing models using cross-sectional regression. *Journal of Finance* 53: 1285–1309.

Jagannathan, R., and Z. Wang. 2002. Empirical evaluation of asset pricing models: A comparison of the SDF and beta methods. *Journal of Finance* 57: 2337–2367.

Jagannathan, R., G. Skoulakis, and Z. Wang. 2002. Generalized method of moments: Applications in finance. *Journal of Business and Economic Statistics* 20: 470–481.

Jensen, M. 1968. The performance of mutual funds in the period 1945–1964. *Journal of Finance* 23: 389–416.

Jobson, J. 1982. A multivariate linear regression test of the arbitrage pricing theory. *Journal of Finance* 37: 1037–1042.

Jobson, J., and B. Korkie. 1982. Potential performance and tests of portfolio efficiency. *Journal of Financial Economics* 10: 433–466.

Jobson, J., and B. Korkie. 1985. Some tests of linear asset pricing with multivariate normality. *Canadian Journal of Administrative Sciences* 2: 114–138.

Kan, R., and G. Zhou. 2001. *Tests of mean-variance spanning*. Working paper, Washington University in St Louis.

Lehman, B., and D. Modest. 1988. The empirical foundations of the arbitrage pricing theory. *Journal of Financial Economics* 21: 213–254.

Merton, R. 1973. An intertemporal capital asset pricing model. *Econometrica* 41: 867–887.

Merton, R. 1981. On market timing and investment performance I: Any equilibrium theory of values for markets forecasts. *Journal of Business* 54: 363–406.

Mitchell, M., and T. Pulvino. 2001. Characteristics of risk and return in risk arbitrage. *Journal of Finance* 56: 2135–2175.

Pastor, L., and R. Stambaugh. 2000. Comparing asset pricing models: An investment perspective. *Journal of Financial Economics* 56: 335–381.

Pastor, L., and R. Stambaugh. 2002. Mutual fund performance and seemingly unrelated assets. *Journal of Financial Economics* 63: 315–349.

Peñaranda, F., and E. Sentana. 2004. *Spanning tests in return and stochastic discount factor mean-variance frontiers: A unifying approach*. Working paper No. 0410. Madrid: CEMFI.

Roll, R., and S. Ross. 1980. An empirical investigation of the arbitrage pricing theory. *Journal of Finance* 35: 1073–1103.

Rosenberg, B., K. Reid, and R. Lanstein. 1984. Persuasive evidence of market inefficiency. *Journal of Portfolio Management* 11: 9–17.

Ross, S. 1976a. The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13: 341–360.

Ross, S. 1976b. Risk, return and arbitrage. In *Risk return in finance*, ed. I. Friend and J. Bicksler. Cambridge, MA: Ballinger.

Shanken, J. 1982. The arbitrage pricing theory: Is it testable? *Journal of Finance* 37: 1129–1240.

Shanken, J. 1985. A multi-beta CAPM or equilibrium APT?: A reply. *Journal of Finance* 40: 1189–1196.

Shanken, J. 1992. On the estimation of beta-pricing models. *Review of Financial Studies* 5: 1–33.

Stambaugh, R. 1983. Arbitrage pricing with information. *Journal of Financial Economics* 12: 357–369.

Wang, Z. 2005. A shrinkage approach to model uncertainty and asset allocation. *Review of Financial Studies* 18: 673–705.

Wang, Z., and X. Zhang. 2005. *Empirical evaluation of asset pricing models: Arbitrage and pricing errors over contingent claims*. Working paper. Federal Reserve Bank of New York.

Arbitrage, Information Theft and Insider Trading

Michael C. Jensen

Abstract

Risk arbitrage involves the purchase of a target firm's shares on the announcement of a merger or tender offer. These transactions provide a risky profit opportunity when the price of the target is below the risk-adjusted expected value of the final takeover price. This article explores the role of arbitragers in the merger and acquisition of firms, and how their role is important to the process and is not to be confused with insider-trading.

Keywords

Takeovers; Mergers and acquisitions; Securities and Exchange Commission; Arbitragers

JEL Classifications

G3; G32; G34

Arbitrage, as the word is used in economics, refers to the simultaneous purchase and sale of an identical commodity. Arbitrage limits the possibilities for the price of a commodity to differ by more than the transactions and transportation costs of moving the good from one place to another. Arbitrage in the financial markets, more accurately known as risk arbitrage because it does not usually involve the simultaneous purchase and sale of the same item, became popular as well as controversial during the heyday of the corporate control market in the 1970s and 1980s. Risk arbitrage in the control market involves the purchase of a target firm's shares (and sometimes the sale of the offering-firm's shares) on the announcement of a merger or tender offer. These transactions provide a risky profit opportunity when the price of the target is below the risk-adjusted expected value of the final takeover price. Sometimes the transactions are executed by speculators who are forecasting the arrival of a takeover offer, or who are acting on rumours of a forthcoming offer.

Arbitragers provide important productive services to investors, and the supply of these services is threatened by the outpouring of protests and legal actions in the wake of the United States Securities and Exchange Commission (SEC) and Justice Department prosecution of insider trading cases. Much of the opposition to arbitrage and the attempt to identify it with insider trading fails to recognize that target shareholder interests are served by a legal rule that allows the producer of privately created information to share that information with others, including arbitragers. There is no economic basis for barring trading on this information so long as the information is legally obtained.

When takeover bids occur, arbitragers provide valuable services for target-firm investors who do not have the time, ability or inclination to gather information on takeover bids for companies in which they hold stock. They help direct resources to their highest-valued use. In doing so, arbitragers provide three critically important services: (1) they help value alternative offers when they occur, including the plans (and implicit offers) of target management, (2) they provide risk-bearing services for investors who do not wish to bear the great uncertainty that occurs between the announcement and final outcome of a takeover bid or restructuring, and (3) they help resolve the collective action or free rider problems of small diffuse shareholders who cannot organize to negotiate directly with competing bidders for the target firm. The arbitragers do this by aggregating large blocks of shares for tender to the highest bidder – sometimes even negotiating the offer price directly with the bidder. Many investors take advantage of the arbitragers' services. This is evidenced in the US by the large fraction of target-company shares in successful takeovers that are frequently held by arbitragers at the completion of the transaction.

The indiscriminate attack on arbitrage and the attempts to define it as insider trading threaten to damage small investors, capital markets and corporations. The threat arises because of the failure to distinguish between two very different situations.

The first situation occurs when an individual steals information from his employer (say an investment bank or a corporation) and/or his employer's client or any other person or organization. Information theft is similar to the theft of any property and should be appropriately prosecuted. Information theft also occurs when corporate managers, acting as the agents of shareholders, produce valuable information about their own company and appropriate its value without permission. Because managers are in a fiduciary relationship with their stockholders, they do not have the right to claim the value of the information unless their contract with the shareholders gives them that right. Under current SEC rules such contracts that allow managers to trade on inside information are prohibited (although as Henry Manne, Dean of the George Mason Law School, pointed out long ago, this prohibition can harm shareholders when such insider trading would be an optimal way to compensate managers). In the absence of such contracts the use of private company information by its executives or employees is theft of shareholder assets and should be prosecuted as theft.

The second situation occurs when an individual produces valuable information about another firm and voluntarily shares it or sells it to others. This sharing of information is no different from any other exchange and should not be prohibited. After expending resources to produce valuable information for themselves about how a target company can be restructured to create value, takeover specialists can rationally decide to share that information with others (including outright sale) prior to releasing it to the public. In this case, trading on such shared information damages no one, and if such sharing is prohibited or discouraged as under current SEC policy, the very investors the SEC seeks to protect (non-insiders) will be harmed.

There are several reasons why takeover specialists would sometimes want to share the value of information they have personally created with others, and in particular with arbitragers. Suppose a specialist has exhausted his or her capital and borrowing power, and does not have enough shares to ensure victory in a hostile offer for control of the target. By sharing his valuable information with arbitragers (who specialize in evaluating proposed deals and betting their own money and that of their investors on the outcomes) the bidder can enlist the arbitragers' help in accumulating enough shares to accomplish the deal. The deals are by no means certain, and the arbitragers will lose in some, as happened in the US market crash of October 1987 when many deals fell apart. Arbitragers are compensated for the valuation, risk-bearing and collective action services they provide by the gains they make from the private information they create or that they receive from the takeover specialist.

The sharing of private information by takeover specialists with arbitragers does not harm other investors. To conclude the opposite, as many have done, assumes that other investors in the target company have a right to claim the value of the information created by the takeover specialist. I have never seen a reasoned argument, either moral or economic, that justifies such a claim. To the contrary, it is generally accepted that someone who paints a picture or builds a house with his or her own resources owns that painting or house and has claim to the value it commands in the marketplace. Such rules encourage productive effort, and provide the economic basis for progress. Application of this principle to entrepreneurial takeover activities implies that the producer of valuable information about the creation of value in target firms should also have claim to the value of that information. Rejection of this principle will reduce the resources devoted to these information production activities, activities that are bringing about the restructuring and enhancement of corporations.

Most important, however, current law fails to comprehend that the interests of target shareholders are served by a legal rule that allows the producer of private information to share that information with others, including arbitragers. The foundations of this proposition are already recognized by the SEC when the legal form of the sharing relationship is formal. For example, T. Boone Pickens, as CEO of Mesa Petroleum, shared with two other individuals, Cyril Wagner and Jack Brown, information Mesa produced

regarding the enormous value that could be created by restructuring Unocal. The relationship between Mesa and Wagner and Brown was formalized in a legal partnership agreement creating Mesa Partners II. In return, Wagner and Brown contributed capital and their own talents and information to the partnership to be used for acquisition of Unocal stock.

Target shareholders are better served by a legal system that makes it possible for takeover specialists to share information with others without a formal pre-offer partnership agreement. The partnership agreement locks suppliers of capital and risk-bearing services, such as Wagner and Brown, into a binding arrangement with the bidder whereas the informal sharing of information with arbitragers does not. This informal sharing of information benefits target shareholders because it leaves the arbitragers in a position to tender their shares to another bidder or the target company itself if either produces a competing bid more valuable than that of the original bidder. One can be sure that the arbitragers, with hundreds of millions of dollars invested in the target's stock, are highly motivated to discover and tender their shares to the highest bidder. In doing so, they help to ensure that the resources of the target firm go to their highest-valued use and that shareholders who hold their shares through the bidding period receive the maximum payment for their shares.

An investor who buys and holds can always be assured of receiving the full value of the information created by potential bidders. Shareholders who voluntarily sell prior to the announcement or outcome of the contest are not damaged, and they gain when they sell at prices higher than would exist in the absence of bidder and/or arbitrager activity. Those investors who voluntarily sell shares to bidders and arbitragers do so because, given their information, they believe the price they receive is higher than the value they place on the firm. Such investors profit at the arbitragers' expense when they sell prior to the failure of a takeover bid. On the other hand they lose some of the gain when they sell prior to the completion of a successful takeover. In the latter case they would be better off if they had waited to sell until after the full information became available to the market. Such investors do not, however, have either a moral or an economic claim to this information or its value, and giving them a legal claim will harm all investors by stifling the production of new information and takeovers. To avoid this damage the current definition of insider trading should be clarified to make clear that the sharing of legally acquired information between creators of valuable information (including takeover specialists) and others (including arbitragers) is legal.

Much confusion is generated by the term insider trading. The notion that all investors should have equal information while executing trades leads to policy recommendations that threaten grave damage to markets, productivity and economic efficiency. Substituting the phrase 'information theft' for insider trading accurately characterizes the subset of information acquisition and trading activities that are economically damaging and therefore should be penalized. Giving or selling information used in securities trading is economically productive as long as that information is not stolen. Abolition of the term 'insider trading' would produce a major improvement in the public, legal and scholarly discussion of these important policy issues.

See Also

- ▶ Corporate Law, Economic Analysis of
- ▶ Credit Rating Agencies
- ▶ Information Sharing among Firms
- ▶ Insider Trading
- ▶ Merger Analysis (United States)

Bibliography

Carlton, D.W., and D.R. Fischel. 1983. The regulation of insider trading. *Stanford Law Review* 35: 857–895.

Easterbrook, F.H. 1981. Insider trading, secret agents, evidentiary privileges, and the production of information. *Supreme Court Review* 11: 309.

This article was first published in The New Palgrave Dictionary of Money and Finance, 1992, edited by John Eatwell, Murray Milgate and Peter Newman and is reproduced here with permission.

Haddock, D.D., and J.R. Macy. 1987. Regulation on demand: A private interest model, with an application to insider trading. *Journal of Law and Economics* 30: 31–52.

Manne, H.G. 1966. *Insider trading and the stock market*. New York: Free Press.

Manne, H.G. 1974. Economic aspects of required disclosure under federal securities laws. In *Wall Street in transition*, ed. H.G. Manne and E. Solomon. New York: New York University Press.

Schotland, R. 1967. Unsafe at any price: A reply to Manne. *Virginia Law Review* 53: 1425.

Arbitration

John T. Dunlop

Arbitration is the process of resolving disputes between two or more parties in which an individual or a board of arbitrators is authorized to appraise the facts and contending positions and to render a decision binding on the parties to the proceeding.

Arbitration is most extensively used in industrial relations, in disputes between labour organizations and managements. The process has been adapted to a widening variety of disputes such as in some landlord–tenant issues, divorce settlements, home or product warranties, in the interpretation of some commercial contracts and even in the settlement of some international questions, as in relative fishing rights between two countries.

Arbitration is said to be *voluntary* when the parties agree voluntarily to enter the process and to be bound by the decision. Arbitration is said to be *compulsory* when the parties are required by law to submit the dispute to a determination and to be bound by the decision. In voluntary arbitration the disputing parties are typically free to frame the question to be resolved, to select the arbitrator or the process of selection, to elect the form of arbitration and to shape the timing and the process. They also typically pay for the arbitration service. Under compulsory arbitration the parties may also have some role in selecting the arbitrator, or in the process of selection, or in influencing features of the process, but they have no choice but to submit to an arbitration procedure often specified in detail in statute.

Arbitration is to be distinguished from mediation, conciliation and fact-finding. While these processes are also widely used to facilitate the resolution of disputes, unlike arbitration there is no authorization to render a decision that is binding on the parties. Mediators typically seek to persuade contending parties to agree, and fact-finders typically make specific recommendations for a voluntary settlement, but they have no authority to issue a binding award. The world of experience does not readily fit neatly into these definitional boxes; arbitration proceedings may involve mediation, and an arbitration award may in fact reflect full agreement of the parties, and the parties may prefer that the arbitrator(s) take responsibility for the ‘award’ before the public and their constituencies. The ‘award’ may in fact be an agreement of the parties or their representatives.

Arbitration is not a single invariant process, since at least in voluntary arbitration the parties have wide latitude to shape its form apart from the selection of the arbitrator(s). Arbitration may be of the last-best-offer variant in which the parties each present to the arbitrator(s) a final position, and the arbitrator(s) is required to select only one or the other proposal. By contrast in conventional arbitration the decision need not adopt either of the contending positions. The parties may also shape the arbitration process by defining the limits on the authority delegated to the arbitrator. The arbitrator may be restricted to the application or interpretation of an agreement or in the remedy the arbitrator may specify. Each of the parties may appoint an arbitrator, or a non-voting assessor to sit with an arbitrator, and they may in turn select the chair. The voting within the board of arbitration may be by majority vote or by the single vote of the chair, materially affecting the outcome in some cases. Thus the parties to the dispute may design the voluntary arbitration process in a wide variety of ways.

In the industrial relations system of the United States the distinction is drawn, as was not drawn historically in England, between issues of right

(questions over the interpretation and application of a collective agreement) and issues of interest (questions concerning the terms of an agreement or issues outside an agreement). This distinction has been fundamental in the United States to the role of the grievance and arbitration procedure, the specified duration of collective agreements and the no-strike no-lockout provisions that limit industrial conflict in the United States.

Historically, in Great Britain, collective agreements had no fixed duration; an agreement could be reopened by either party on specified notice or with a specified change in some exterior event such as prices or trade. A dispute between a labour union and a management could equally be over an interpretation of an existing agreement or over a proposed change in the agreement itself. Such a distinction was not made. A strike or lockout could as readily be used as a tool to reach agreement in either case. Arbitration had no special role except as might be agreed upon in the particular dispute.

In the United States, in contrast, disputes over the interpretation or application of the agreement came voluntarily in many industries historically to be referred to standing arbitration tribunals or ad hoc arbitrators. The strike or lockout was precluded during the term of the agreement. Arbitration of issues of interpretation and application was the quid pro quo for both parties for giving up resort to economic force for a limited period. A no-strike, no-lockout clause was not possible in a labour agreement of any extended duration without arbitration to resolve grievances over the interpretation or application of that agreement.

This role for grievance arbitration in the United States long antedates the labour legislation of the 1930s or the 1960 decision of the Supreme Court in the *Steelworkers' Trilogy* that established a limited role for the courts to review arbitration awards. Thus the Anthracite Board of Conciliation, set up in 1903, and an early 'intermittent' umpire, Judge George A. Gray, established the rule that 'the Board could not write the law, but could only interpret it.' The clothing industries early used impartial umpires to settle disputes over piece rates and other terms of the agreement, but they also had a role in helping the parties by mediation and at times by arbitration to settle the terms of collective agreements. As industrial plants were organized on an industrial basis, the principle was carried over into these collective agreements with each collective bargaining relationship designing its own grievance arbitration procedures.

Beyond grievance arbitration in the United States, which encompasses the largest part of industrial relations arbitration, there are significant instances of arbitration over the terms of collective agreements, particularly in the public sector in some states.

There have been at least two contending views as to the nature of the arbitration process and the considerations that lead to the decision of the arbitrator. One view is that arbitrators act like judges are supposed to act: they weigh the facts and arguments against the standards and precedents urged by the parties to the conflict and render a decision with an articulated opinion. Another view is that arbitrators are primarily concerned to achieve a mutually acceptable solution, a position that the parties themselves would have achieved in their bargaining or administration had it not been frustrated and fallen short of full agreement. There are, no doubt, pairs of parties and arbitrators that follow each perspective; others fall in between. In their bargaining the parties seek to shape the process and the choice of arbitrators accordingly.

See Also

- ▶ Bargaining
- ▶ Industrial Relations
- ▶ Trade Unions

Bibliography

Aaron, B. 1983. No labour courts, little arbitration: What's wrong with that? In *Comparative industrial relations: A trans-atlantic dialogue.* Washington, DC: Bureau of National Affairs. 56–70.

Donovan, Lord. (Chairman). 1968. *Royal commission on trade unions and employer associations*. Cmd 3623. London: HMSO.

Dunlop, J.T. 1984. *Dispute resolution, negotiation and consensus building*. Dover: Auburn House.
Elkouri, F., and E.A. Elkouri. 1985. *How arbitration works*, 4th ed. Washington, DC: BNA Books.
Kennedy, T. 1948. *Effective labor arbitration: The impartial chairmanship of the full-fashioned hosiery industry*. Philadelphia: University of Pennsylvania.
Lester, R.A. 1984. *Labor arbitration in state and local government*. Princeton: Industrial Relations Section.
Lowell, J.S. 1893. *Industrial arbitration and conciliation*. New York: Putnam's Sons.
National Academy of Arbitrations. 1943 onwards. *Proceedings of the annual meetings*. Washington, DC: BNA.
Stevens, C.M. 1966. Is compulsory arbitration compatible with bargaining? *Industrial Relations Review* 5(2): 38–52.
Suffern, A.E. 1915. *Conciliation and arbitration in the coal industry of America*. Boston: Houghton-Mifflin.
Walker, K.F. 1970. *Australian industrial relations systems*. Cambridge, MA: Harvard University Press.

ARCH Models

Oliver B. Linton

Abstract

The ARCH model and its many generalizations are very important in analysing discrete time financial data. We review the properties of the original model and discuss many of the subsequent developments.

Keywords

ARCH models; ARMA models; Estimation; Exponentially weighted moving average model; Factor models; GARCH models; Generalized error distribution; Heteroskedasticity; IGARCH models; Linear models; Long memory models; Multivariate models; News impact curve; Nonparametric models; Semiparametric models; Stationarity; Time series analysis; Unit roots

JEL Classifications

C22

Introduction of Model and Basic Properties

The key properties of financial time series appear to be that: (*a*) marginal distributions have heavy tails and thin centres (leptokurtosis); (*b*) the scale appears to change over time; (*c*) return series appear to be almost uncorrelated over time but to be dependent through higher moments (see Mandelbrot 1963; Fama 1965). Linear models like the autoregressive moving average (ARMA) class cannot capture well all these phenomena, since they only really address the conditional mean $\mu_t = E(y_t|\, y_{t-1}, \ldots)$ and in a rather limited way. This motivates the consideration of nonlinear models. For a discrete time stochastic process y_t, the conditional variance $\sigma_t^2 = \mathrm{var}(y_t|\, y_{t-1}, \ldots)$ of the process is a natural measure of risk for an investor at time $t-1$. Empirically it appears to change over time and so it is important to have a model for it. Engle (1982) introduced the autoregressive conditional heteroskedasticity (ARCH) model

$$\sigma_t^2 = \omega + \gamma y_{t-1}^2, t = 0, \pm 1, \ldots,$$

where for simplicity we rewrite $y_t \mapsto y_t - \mu_t$ and suppose that the process started in the infinite past. This model makes σ_t^2 vary over time depending on the realization of past squared returns. For σ_t^2 to be a valid conditional variance it is necessary that $\omega > 0$ and $\gamma \geq 0$, in which case $\sigma_t^2 > 0$ for all t. Suppose also that $y_t = \varepsilon_t \sigma_t$ with ε_t i.i.d. mean zero and variance one. Provided $\gamma < 1$, the process y_t is weakly (covariance) stationary and has finite unconditional variance $\sigma^2 = E(\sigma_t^2) = E(y_t^2) = \omega/(1-\gamma)$. This can be proven rigorously under a variety of assumptions on the initialization of the process (see Nelson 1990). The meaning of this is that the process fluctuates about the long-run value σ^2 and forecasts converge to this value as the forecast horizon lengthens.

The ARCH process is dynamic like ARMA models and indeed we can write the process as an AR(1) in y_t^2, that is,

$$y_t^2 = \omega + \gamma y_{t-1}^2 + \eta_t,$$

where $\eta_t = y_t^2 - \sigma_t^2 = \sigma_t^2(\varepsilon_t^2 - 1)$ is a mean zero, uncorrelated sequence, that is heteroskedastic. Therefore, we generally have dependence in σ_t^2, y_t^2, and because of the parameter restrictions, positive dependence that is, cov $\left(\sigma_t^2, \sigma_{t-j}^2\right) > 0$ and cov$\left(y_t^2, y_{t-j}^2\right) > 0$. As far as the second order properties (that is, the covariance function) of the process y_t^2, this is identical to that of an AR(1) process. However, it should be remembered that y_t^2 is heteroskedastic itself and that the form of the heteroskedasticity has to be particularly extreme since y_t^2 is kept non-negative.

One feature of linear models like the ARMA class is that the marginal distribution of the variable is normally distributed whenever the shocks are i.i.d. normally distributed. This is not the case for the ARCH class of processes. Specifically, the marginal distribution of y_t will be heavy tailed even if $\varepsilon_t = (y_t - \mu_t)/\sigma_t$ is standard normal. Suppose ε_t is standard normal (and the process is weakly stationary), then the excess kurtosis of y_t is $\kappa_4 = 6\gamma^2/(1 - 3\gamma^2) \geq 0$ provided $\gamma^2 < 1/3$. If $\gamma \geq 1/3^{1/2}$, then $E\left(y_t^4\right) = \infty$. For leptokurtic ε_t, the restriction on γ for finite fourth moment is even more severe. Although the ARCH(1) model implies heavy tails and volatility clustering, it does not in practice generate enough of either. The constraint on γ for finite fourth moment severely restricts the amount of persistence; it is an undesirable feature that the same parameter controls both persistence and heavy tailedness, although if one allows non-normal distributions for e_t, this link is broken on one side at least.

The extension to the ARCH(p) process with p lags, while more flexible, becomes very complicated to estimate without restrictions on the coefficients. Bollerslev (1986) introduced the GARCH(p,q) process

$$\sigma_t^2 = \omega + \sum_{k=1}^{p} \beta_k \sigma_{t-k}^2 + \sum_{j=1}^{q} \gamma_j \left(y_{t-j} - \mu_{t-j}\right)^2,$$

whose $p = 1$, $q = 1$ GARCH(1,1) special case contains only three parameters and usually does a better job than an unrestricted ARCH(12), say, according to a variety of statistical criteria. The GARCH(1,1) process is probably still the most widely used model. As with the ARCH process one needs restrictions on the parameters to make sure that σ_t^2 is positive with probability one. For the GARCH(1,1) it is necessary that $\gamma, \beta \geq 0$ and $\omega > 0$. Interestingly, for higher order processes it is not necessary that ω, γ_j, $\beta_j \geq 0$ for all j: see Nelson and Cao (1992). For example, in GARCH (1,2) the conditions are that β, $\gamma_1 \geq 0$ and $\beta\gamma_1 + \gamma_2 \geq 0$. Provided $\sum_{k=1}^{p} \beta_k + \sum_{j=1}^{q} \gamma_j < 1$, the process y_t is weakly stationary and has finite unconditional variance

$$\sigma^2 = E\left(\sigma_t^2\right) = \frac{\omega}{1 - \sum_{k=1}^{p} \beta_k - \sum_{j=1}^{q} \gamma_j}.$$

As for the ARCH process, the series y_t has higher kurtosis than ε_t.

Drost and Nijman (1993) provide an important classification of ARCH models according to the precise properties required of the error terms. The strong GARCH process is where

$$\varepsilon_t = \frac{y_t - \mu_t}{\sigma_t} \text{ i.i.d.} \quad E(\varepsilon_t) = 0 \quad \text{and} \quad E\left(\varepsilon_t^2\right) = 1.$$

It is generally this case that has been investigated in the literature. It is a very strong assumption by the standards of most modern econometrics, where usually only conditional moment restrictions are imposed, but is a complete specification that is useful for deriving properties like stationarity. The strong Gaussian case is where ε_t is additionally normally distributed. The semi-strong GARCH process is where

$$E[\varepsilon_t | y_{t-1}, y_{t-2}, \ldots] = 0 \quad \text{and} \quad E\left[\varepsilon_t^2 | y_{t-1}, y_{t-2}, \ldots\right] = 1.$$

These assumptions are weaker and turn out to be sufficient in many cases for consistent estimation. They are quite weak assumptions and restrict only the conditional mean and conditional variance of the process, allowing a variety of behaviour in the potentially time varying distribution of ε_t. Drost and Nijman (1993) show that conventional strong and semi-strong GARCH processes are not closed under temporal aggregation, meaning that if a process is GARCH at the daily

frequency that the weakly or monthly data may not be GARCH, either weak or strong.

Strong Stationarity and Mixing

Consider the GARCH(1,1) process

$$y_t = \sigma_t \varepsilon_t, \quad \sigma_t^2 = \omega + \beta \sigma_{t-1}^2 + \gamma y_{t-1}^2$$

with ε_t i.i.d. and $\omega > 0$ and $\beta, \gamma \geq 0$. A sufficient condition for strong stationarity is that $E\left[\ln(\beta + \gamma \varepsilon_t^2\right] < 0$ (see Nelson 1990). If additionally, $E(\varepsilon_t) = 0$ and $\text{var}(\varepsilon_t) = 1$, then the necessary and sufficient condition for weak stationarity is that $\beta + \gamma < 1$. By Jensen's inequality $E\left[\ln(\beta + \gamma \varepsilon_t^2\right] < \ln\ E\left[(\beta + \gamma \varepsilon_t^2\right] = \ln(\beta + \gamma)$, so it can be that $E\left[\ln\left(\beta + \gamma \varepsilon_t^2\right)\right] < 0$ even when $\beta + \gamma \geq 1$, that is, there are strongly stationary processes that are not weakly stationary.

There are many measures of dependence in time series. Mixingness is the property that dependence dies out with horizon. It can be measured in different ways: covariance mixing, strong mixing, and beta mixing are the main concepts. A stationary sequence $\{X_t,\ t = 0, \pm 1, \ldots\}$ is said to be covariance mixing if $cov(X_t, X_{t+k}) \rightarrow 0$ as $k \rightarrow \infty$. A stationary sequence $\{X_t, t = 0, \pm 1, \ldots\}$ is said to be strong mixing (α-mixing) if

$$\alpha(k) = \sup_{A \in \mathcal{F}_{-\infty}^n B \in \mathcal{F}_{n+k}^\infty} |P(AB) - P(A)P(B)| \rightarrow 0$$

as $k \rightarrow \infty$, where $\mathcal{F}_{-\infty}^n$ and $\mathcal{F}_{n+k}^\infty$ are two σ-fields generated by $\{X_t,\ t \leq n\}$ and $\{X_t,\ t \geq n + k\}$, respectively. We call $\alpha(\cdot)$ the mixing coefficient. A stationary sequence $\{X_t, t = 0, \pm 1, \ldots\}$ is said to be β-mixing if

$$\beta(k) = \sup_{A \in \mathcal{F}_{-\infty}^n B \in \mathcal{F}_{n+k}^\infty} |P(AB) - P(A)P(B)| \rightarrow 0$$

as $k \rightarrow \infty$ .We call $\beta(\cdot)$ the mixing coefficient. We have $2\alpha\ (k) \leq \beta(k)$. The covariance mixing property is only well defined for weakly stationary processes, so it is natural here to work with the more general notions of α and β mixing. A sufficient condition that a GARCH(1,1) process is β-mixing with exponential decay is that it is weakly stationary, Carrasco and Chen (2002), but this is not necessary. More recently it has been shown that IGARCH is strong mixing under some conditions (see Meitz and Saikkonen 2004). One problem is that when you combine a GARCH process with other processes for the mean, the mixingness is not preserved and has still to be established. The weaker concept of near epoch dependence can be established, though in quite a general class of models (Hansen 1991). Why does mixing matter? It is a key property that allows one to learn from the data through the law of large numbers and central limit theorems.

IGARCH Models

In practice, estimated GARCH parameters lie close to the boundary of the weakly stationary region. This prompts consideration of the process where $\sum_{k=1}^p \beta_k + \sum_{j=1}^q \gamma_j = 1$, which is called the integrated GARCH or IGARCH. In this case, the process y_t with i.i.d. Gaussian innovations is strongly stationary but not covariance stationary, since the unconditional variance is infinite (although the conditional variance is finite with probability 1). This is in contrast to linear unit root processes in which the process is neither weakly nor strongly stationary and these two notions coincide. Also, in contrast to the linear case, differencing does not induce weak stationarity, that is, $y_t^2 - y_{t-1}^2$ is not weakly stationary (although its mean is constant over time).

The exponentially weighted moving average model (sometimes called the J.P. Morgan model) is a variant on the IGARCH model in which there is no intercept ω and a unit root:

$$y_t = \sigma_t \varepsilon_t, \sigma_t^2 = \beta \sigma_{t-1}^2 + (1 - \beta) y_{t-1}^2.$$

It is a very simple process with only one parameter and is widely used by practitioners, with particular values of the parameter β. Write σ_t^2

$= \sigma_{t-1}^2\left[\beta + (1-\beta)\varepsilon_{t-1}^2\right]$, so that $\ln\sigma_t^2$ is a random walk, that is,

$$\begin{aligned}\ln\sigma_t^2 &= \ln\sigma_{t-1}^2 + \eta_{t-1}, \quad \eta_{t-1} \\ &= \ln\left(\beta + (1-\beta)\varepsilon_{t-1}^2\right),\end{aligned}$$

and hence is not strongly stationary. On the other hand, the process y_t is informally weakly stationary since $E\left[y_t^2 \middle| \mathcal{F}_{-\infty}^0\right] = E\left[\sigma_t^2 \middle| \mathcal{F}_{-\infty}^0\right] = \prod_{s=1}^t E\left[\beta + (1-\beta)\varepsilon_s^2\right]\sigma_0^2 = \sigma_0^2$ for all t. The properties of this process depend on the moments of η_{t-1}. If $E[\eta_{t-1}] > 0$, then $\ln\sigma_t^2 \to \infty$ with probability 1. If $E[\eta_{t-1}] < 0$, then $\sigma_t^2 \to -\infty$ with probability 1 as $t \to \infty$ and so $\sigma_t^2 \to 0$ with probability 1. If $E[\eta_{t-1}] = 0$, then $\ln\sigma_t^2$ is a driftless random walk and the process just wanders everywhere. If we assume $E\left[\varepsilon_t^2\right] = 1$, then by Jensen's inequality $E[\eta_{t-1}] < 0$, and the process $\sigma_t^2 \to 0$ with probability 1 as $t \to \infty$ whatever the initialization. Thus the process is essentially degenerate and is not plausible, despite being widely used.

Functional Form

The news impact curve is the relationship between σ_t^2 and $y_{t-1} = y$ holding past values σ_{t-1}^2 constant at some level σ^2. This is an important relationship that describes how new information affects volatility. For the GARCH process, the news impact curve is

$$m\left(y, \sigma^2\right) = \omega + \gamma y^2 + \beta\sigma^2.$$

It is separable in σ^2, it is an even function of news y, $m(y, \sigma^2) = m(-y, \sigma^2)$, and it is a quadratic function of y. The symmetry property implies that $\operatorname{cov}\left(y_t^2, y_{t-j}\right) = 0$ for symmetric about zero ε_t.

The GARCH process does not allow 'leverage effects' or asymmetric news impact curves. Because of limited liability, we might expect that negative and positive shocks have different effects on volatility. Nelson (1991) introduced the exponential GARCH model. Let $h_t = \log\sigma_t^2$ of and let

$$h_t = \omega + \sum_{j=1}^{p}\gamma_j\left[\theta\varepsilon_{t-j} + \delta\left|\varepsilon_{t-j}\right|\right] + \sum_{k=1}^{q}\beta_k h_{t-k},$$

where $\varepsilon_t = (y_t - \mu_t)/\sigma_t$ is i.i.d. with mean zero and variance one. Nelson's paper contains four innovations. First, it models the log, not the level. Therefore there are no parameter restrictions to ensure that $\sigma_t^2 \geq 0$. Second, it allows asymmetric effect of past shocks ε_{t-j} on current volatility, that is, the news impact curve is allowed to be asymmetric. For example, $\operatorname{cov}\left(y_t^2, y_{t-j}\right) \neq 0$ even when ε_t is symmetric about zero. Third, it makes the innovations ε_t i.i.d. It follows that h_t is a linear process so that strong and weak stationarity coincide where they ought to (for h_t anyway). On the other hand estimation and forecasting is quite tricky because of the repeated exponential/logarithmic transformations involved. The final innovation was to allow heavy tailed innovations based on the so-called generalized error distribution (GED) that nests the Gaussian as a special case.

An alternative approach to allowing asymmetric news impact curve is the Glosten et al. (1993) model

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \gamma y_{t-1}^2 + \delta y_{t-1}^2 1(y_{t-1} < 0).$$

In this case, the news impact curve is asymmetric but still has quadratic tails. It is a simple enough modification, that it has similar probabilistic properties to the GARCH(1,1) process. There are many other variations on the basic GARCH model, too many to list here, but the interested reader can find a fuller description in the survey paper of Bollerslev et al. (1994).

One might expect that risk and return should be related: see Merton (1973) for an example. The GARCH-in-Mean process captures this idea. This process is

$$y_t = g\left(\sigma_t^2; b\right) + \varepsilon_t\sigma_t,$$

for various functional forms of g, for example, linear and log-linear and for some given GARCH specification of σ_t^2. Engle et al. (1987) used this model on interest rate data (see also Pagan and Hong 1991). Here, b are parameters to be estimated along with the parameters of the

error variance. Some authors find small but significant effects.

Estimation

The standard approach to estimation of these models has been through estimation of the (conditional) Gaussian quasi-likelihood criterion

$$L_T(\theta) = \sum_{t=1}^{T} \ell_t(\theta) = -\sum_{t=1}^{T} \log\sigma_t^2(\theta) - \frac{1}{2}\sum_{t=1}^{T}\left(\frac{y_t - \mu_t(\theta)}{\sigma_t(\theta)}\right)^2,$$

where $\sigma_t^2(\theta)$ and perhaps $\mu_t(\theta)$ are built up by recursions from some starting values. There are several possibilities regarding starting values: (a) $\sigma_0^2(\theta) = \omega/1 - \beta - \gamma$, (b) $\sigma_0^2(\theta) = T^{-1}\sum_{t=1}^{T} y_t^2$, and (c) $\sigma_0^2(\theta) = y_1^2$. Approach (a) imposes weak stationarity and would not be appropriate were IGARCH to be thought plausible, while value (b) sort of requires weak stationarity for the asymptotic properties to follow through. The likelihood function is maximized with respect to the parameter values usually using some derivative-based algorithm like BHHH and sometimes imposing inequality restrictions (like those required for $\sigma_t^2 \geq 0$ with probability 1 or for σ_t^2 to be weakly stationary) and sometimes not.

The (quasi) MLE (QMLE) can be expected to be consistent provided only the conditional mean and the conditional variance are correctly specified (Bollerslev and Wooldridge 1992), that is, semi-strong not strong GARCH is required and conditional normality is certainly not required. This is true because the score function $\partial\ell_t(\theta_0)/\partial\theta$ is a martingale difference sequence. Robust standard errors can be constructed in the usual way

$$\left[\frac{\partial\ell_T\left(\hat{\theta}\right)}{\partial\theta\partial\theta^{\top}}\right]^{-1}\left[\sum_{t=1}^{T}\frac{\partial\ell_t}{\partial\theta}\frac{\partial\ell_t}{\partial\theta^{\top}}\left(\hat{\theta}\right)\right]\left[\frac{\partial\ell_T\left(\hat{\theta}\right)}{\partial\theta\partial\theta^{\top}}\right]^{-1}, \quad (1)$$

although the default option in many software packages is to compute standard errors as if Gaussianity held.

The distribution theory is difficult to establish from primitive conditions even for simple models. There is one important point about these asymptotics – that one does not need moments on y_t (for example, one does not need weak stationarity). Lumsdaine (1996) established consistency and asymptotic normality allowing the IGARCH case but under strong stationarity and symmetric unimodal i.i.d. ε_t with $E\left[\varepsilon_t^{32}\right] < \infty$. Lee and Hansen (1994) proved the same result under weaker conditional moment conditions and allows for semi-strong processes with some higher-level assumptions. Jensen and Rahbek (2004) established consistency and asymptotic normality of the QMLE in strong GARCH model without strict stationarity. Hall and Yao (2003) assume weak stationarity and show that if $E\left[\varepsilon_t^4\right] < \infty$ the asymptotic normality holds, but also establish limiting behaviour (non-normal) under weaker moment conditions. No results have yet been published for consistent and asymptotically normality of EGARCH from primitive conditions, although simulation evidence does suggest normality is a good approximation in large samples.

Typically, one finds small intercepts and a large parameter on the lagged dependent volatility; see Lumsdaine (1995) and Brooks et al. (2001) for simulation evidence. These two parameter estimates are often highly correlated. Engle and Sheppard (2001) suggested a method they called target variance to obviate the computational difficulties sometimes encountered in estimating GARCH models. For a weakly stationary GARCH(1,1) process we have $E\left(y_t^2\right) = \omega/(1 - \beta - \gamma)$ so that $\omega = E\left(y_t^2\right)(1 - \beta - \gamma)$. They suggest replacing $E\left(y_t^2\right)$ by $\sum_{t=1}^{T} y_t^2/T$ in the likelihood so that one only has two parameters to chose. This results in a much more stable performance of most algorithms. The downside with this approach is that distribution theory is much more complicated due to the lack of martingale property, and in particular one needs to use Newey–West standard errors.

It is quite common now to estimate GARCH models using different objective functions suggested by alternative specifications of the error distribution like the t or the GED distribution that Nelson (1991) favoured. These objective functions often have additional parameters such as the degrees of freedom that have to be computed. They lead to greater efficiency when the chosen specification is correct, but otherwise can lead to inconsistency, as was shown by Newey and Steigerwald (1997).

Long Memory

The GARCH(1,1) process $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \gamma y_{t-1}^2$ is of the form

$$\sigma_t^2 = c_0 + \sum_{j=1}^{\infty} c_j y_{t-j}^2 \qquad (2)$$

for constants c_j satisfying $c_j = \gamma\beta^{j-1}$, provided the process is weakly stationary, which requires $\gamma + \beta < 1$. These coefficients decay very rapidly so the actual amount of memory is quite limited. There is some empirical evidence on the autocorrelation function of y_t^2 for high frequency data that suggests a slower decay rate than would be implied by these coefficients. Long memory models essentially are of the form (2) but with slower decay rates. For example, suppose that $c_j = j^{-\theta}$ for some $\theta > 0$. The coefficients satisfy $\sum_{j=1}^{\infty} c_j^2 < \infty$ provided $\theta > 1/2$. Fractional integration (FIGARCH) leads to such an expansion. There is a single parameter called d that determines the memory properties of the series, and

$$(1-L)^d \sigma_t^2 = \omega + \gamma\sigma_{t-1}^2\left(\varepsilon_{t-1}^2 - 1\right),$$

where $(1-L)^d$ denotes the fractional differencing operator. When $d = 1$ we have the standard IGARCH model. For $d \neq 1$ we can define the binomial expansion of $(1-L)^{-d}$ in the form given above. See Robinson (1991) and Bollerslev and Mikkelson (1996) for models and evidence of long memory. The evidence for long memory is often based on sample autocovariances of y_t^2, and this may be questionable due to a paper of Mikosch and Stărică (2000).

Multivariate Models

In practice we observe many closely related series, and so it may be important to model their behaviour jointly. Define the conditional covariance matrix

$$\Sigma_t = E\left(y_t y_t^{\top} \mid \mathcal{F}_{-\infty}^{t-1}\right)$$

for some $n \times 1$ vector of mean zero series y_t. Bollerslev, Engle and Wooldridge (1988) introduced the most general generalization of the univariate GARCH(1,1) process

$$h_t = \mathrm{vech}(\Sigma_t) = A + Bh_{t-1} + C\mathrm{vech}\left(y_{t-1}y_{t-1}^{\top}\right),$$

where A is an $n(n+1)/2 \times 1$ vector, while B, C are $n(n+1)/2 \times n(n+1)/2$ matrices. In practice, there are too many parameters. Also, the restrictions on the parameters to ensure that Σ_t is positive definite are very complicated in this formulation. For weak stationarity one requires that the matrix $I - B - C$ is nonsingular and positive definite in which case the unconditional variance matrix is unvech $((I - B - C)^{-1}A)$. The conditions for strong stationarity are rather complicated to state.

The so-called BEKK model is a special case that addresses these issues. It is of the form

$$\Sigma_t = AA^{\top} + B\Sigma_{t-1}B^{\top} + Cy_{t-1}y_{t-1}^{\top}C^{T}$$

for $n \times n$ matrices A, B, C. This gives a big reduction in number of parameters and imposes symmetry and positive definiteness automatically. There are still many parameters that have to be estimated simultaneously, of the order n^2, and this limits the applicability and interpretability of this model.

Bollerslev (1990) introduced the constant conditional covariance (CCC) model, which greatly reduces the parameter explosion issue. This involves standard univariate dynamic models for each of the conditional variances and a constant correlation assumption, that is,

$$\Sigma_t = D_t R D_t, D_t = \text{diag}\{\sigma_{it}\} \quad (3)$$

$$\sigma_{it}^2 = \omega_i + \beta_i \sigma_{i,t-1}^2 + \gamma_i y_{i,t-1}^2 \quad (4)$$

and $R = (Rij)$ is a time invariant matrix

$$R_{ij} = \frac{E\left[\varepsilon_{it}\varepsilon_{ji}\right]}{\left(E\left[\varepsilon_{it}^2\right]E\left[\varepsilon_{jt}^2\right]\right)^{1/2}} = E\left[\varepsilon_{it}\varepsilon_{jt}\right],$$

where $\varepsilon_{it} = y_{it}/\sigma_{it}$ The values R_{ij} are restricted to lie in [−1,1] and the matrix R is symmetric and positive definite but otherwise unrestricted. This model generates time varying conditional covariances, but the dynamics are all driven by the conditional variances as the correlations are constant. The estimation of R is quite straightforward: use the sample correlation matrix of the standardized residuals $\varepsilon_{it} = y_{it}/\hat{\sigma}_{it}$. The estimated matrix R is guaranteed to be symmetric and positive definite because it is a correlation matrix and consequently the estimated Σ_t shares these properties.

Engle and Sheppard (2001) introduced the dynamic conditional covariance (DCC) model where we replace in (3) and (4)

$$R_{ij,t} = \frac{q_{ij,t}}{\left(q_{ij,t}q_{jj,t}\right)^{1/2}}$$
$$q_{ij,t} = c_{ij} + b_{ij}q_{ij,t-1} + a_{ij}\varepsilon_{i,t-1}\varepsilon_{j,t-1}.$$

If we assume also that $a_{ij} = a$, $b_{ij} = b$, and $c_{ij} = c$ for all $i \neq j$ one can show that the resulting covariance matrix Σ_t is guaranteed to be symmetric and positive definite. This model allows slightly more flexibility in allowing the correlations to vary over time, but because of the need to impose positive definiteness it still imposes common dynamics on the correlations, which may be too restrictive.

The approach that brings the most flexible dimensionality reduction is based on the ideas of factor analysis. Suppose that for $y_t \in \mathbb{R}^n, f_t \in \mathbb{R}^k$:

$$y_t = Cf_t + u_t \quad (5)$$

$$\begin{pmatrix} f_1 \\ u_t \end{pmatrix} | I_{t-1} \sim \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Lambda_t & 0 \\ 0 & \Gamma \end{pmatrix}, \quad (6)$$

where $Y_{t-1} = \{y_{t-1}, \ldots\}$ is the observed information and $I_t = \{y_t, f_t, y_{t-1}, f_{t-1}, \ldots\}$ contains both observed series and the latent factors $F_{t-1} = \{f_t, f_{t-1}, \ldots\}$. Suppose that rank$(C) = k$ and that Λ_t is a $k \times k$ positive definite time varying matrix. It follows that $y_t | I_{t-1} \sim 0$, $C\Lambda_t C^\top + \Gamma$ (Sentana 1998). The implied Σ_t is of reduced rank and depends on only order nK (time-varying associated) parameters so there is a big reduction in dimensionality. This model includes as a special case the Diebold and Nerlove (1989) model where Γ, Λ_t are diagonal and $\lambda_{jjt} = var\left[f_{jt} | I_{t-1}\right] = \omega_j + \beta_j \lambda_{jj,t-1} + \gamma_j f_{j,t-1}^2$, in which case $\lambda_{jji} \notin Y_{t-1}$. This process is closed under block marginalization – that is, subsets of y_t do not have the same structure. Estimation is complicated by the latent variables. This framework also includes the Engle et al. (1990) factor GARCH model $\Sigma_t = \Sigma_0 + \sum_{k=1}^K \delta_k \delta_k^\top \sigma_{kt}^2$, where $K < n$, and σ_{kt}^2 is the conditional variance of a certain portfolio k, with time invariant weights σ_k, that is, $y_{kt}^p = \alpha_k^\top y_t$ with $\alpha_k^\top \ i = 1$.

They assume also that σ_{kt}^2 are standard univariate GARCH(1,1) processes, that is, for some parameters $(\omega_k, \beta_k, \gamma_k), \sigma_{kt}^2 = \omega_k + \beta_k \sigma_{k,t-1}^2 + \gamma_k \left(\gamma_k^\top y_{t-1}\right)^2$. This model is written in terms of observables and consequently its estimation is somewhat easier, but it suffers from the fact that it is not closed under block marginalization – that is, subsets of y_t do not have the same structure. Sentana (1998) shows how it is nested in the general model (5) and (6).

Nonparametric and Semiparametric Models

There have been a number of contributions to ARCH modelling from the nonparametric or semiparametric point of view; see Hafner (1998) for an overview. Engle and González-Rivera (1991) suggested treating the error distribution in a GARCH process nonparametrically, that is,

$$y_t = \mu_t + \varepsilon_t \sigma_t \sigma_t^2$$
$$= \omega + \beta \sigma_{t-1}^2 + \gamma (y_{t-1} - \mu_{t-1})^2,$$

where μ_t depends on observed covariates and parameters, while ε_t is i.i.d. with density f that is not restricted in shape. This is motivated by the great deal of evidence that the density of the standardized residuals $\varepsilon_t = (y_t - \mu_t)/\sigma_t$ is non-Gaussian. They proposed an estimation algorithm that involved estimating f from the data. Linton (1993) and Drost and Klaassen (1997) have shown that one can achieve significant efficiency improvements depending on the shape of the error density.

An alternative line of research has been to treat the functional form of $\sigma_t^2(y_{t-1}, y_{t-2}, \ldots)$ nonparametrically. In particular, suppose that

$$\sigma_t^2 = g(y_{t-1}, \ldots, y_{t-p})$$

for some unknown function g and fixed lag length p. This allows for a general shape to the news impact curve and nests all the usual parametric ARCH processes. See Pagan and Hong (1991) and Härdle and Tsybakov (1997) for some applications. This model is somewhat limited in the dependence it allows in comparison with the GARCH(1,1) process, which is a function of all past $y's$. Also, the curse of dimensionality means that the usual estimation methods do not work well in practice for large p, that is, $p > 4$.

One compromise approach to avoiding the curse of dimensionality is to use additive models, whence

$$\sigma_t^2 = \sum_{j=1}^{p} g_j\left(y_{t-j}\right) \tag{7}$$

for some unknown functions g_j. The functions g_j are allowed to be of general functional form but only depend on y_{t-j}. This class of processes nests many parametric ARCH models. The functions g_j can be estimated by kernel regression techniques (see Masry and Tjøstheim 1995). Yang et al. (1999) proposed an alternative nonlinear ARCH model in which the conditional mean is again additive, but the volatility is multiplicative $\sigma_t^2 = c_v \prod_{j=1}^{d} \sigma_j^2\left(y_{t-j}\right)$. Kim and Linton (2004) generalize this model to allow for arbitrary, but known, transformations, that is, $G(\sigma_t^2) = c_v + \sum_{j=1}^{d} \sigma_j^2\left(y_{t-j}\right)$, where $G(.)$ is known function like log or level. Linton and Mammen (2005) considered the case where $\sigma_t^2 = \sum_{j=0}^{\infty} \beta^{j-1} g\left(y_{t-j}\right)$, which nests the GARCH(1,1) process when $g(y) = u + \gamma y^2$.

One final semiparametric approach has been to model the coefficients of a GARCH process as changing over time, thus

$$\sigma_t^2 = \omega(x_{tT}) + \beta(x_{tT})\sigma_{t-1}^2 + \gamma(x_{tT})(y_{t-1} - \mu_{t-1})^2,$$

where ω, β, and γ are smooth functions of a variable x_{tT}, for example, $x_{tT} = t/T$. This class of processes is non-stationary but can be viewed as locally stationary along the lines of Dahlhaus (1997).

See Also

- Continuous and Discrete Time Models
- Factor Models
- Finance
- Local Regression Models
- Martingales
- Time Series Analysis

Acknowledgment The author would like to thank the Economic and Social Science Research Council of the United Kingdom for financial support through a research fellowship.

Bibliography

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*: *307*–327.

Bollerslev, T. 1990. Modelling the coherence in short-run nominal exchange rates: A multivariate generalized autoregressive conditional heteroskedasticity. *Review of Economics and Statistics* 72: 498–505.

Bollerslev, T., R.F. Engle, and J.M. Wooldridge. 1988. A capital asset pricing model with time varying covariances. *Journal of Political Economy* 96: 116–131.

Bollerslev, T., and H.O. Mikkelson. 1996. Modelling and pricing long memory in stock market volatility. *Journal of Econometrics* 73: 151–84, 498–505.

Bollerslev, T., and J.M. Wooldridge. 1992. Quasi maximum likelihood estimation and inference in dynamic models with time varying covariances. *Econometric Reviews* 11: 143–172.

Bollerslev, T., R.Y. Chou, and K. Kroner. 1992. ARCH modelling in finance. *Journal of Econometrics* 52: 5–59.

Bollerslev, T., R.F. Engle, and D. Nelson. 1994. ARCH models. In *The handbook of econometrics*, ed. D.F. McFadden and R.F. Engle III, Vol. 4. Amsterdam: North-Holland.

Brooks, C., S.P. Burke, and G. Persand. 2001. Benchmarks and the accuracy of GARCH model estimation. *International Journal of Forecasting* 17: 45–56.

Carrasco, M., and X. Chen. 2002. Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* 18: 17–39.

Dahlhaus, R. 1997. Fitting time series models to nonstationary processes. *Annals of Statistics* 25: 1–37.

Diebold, F.S., and M. Nerlove. 1989. The dynamics of exchange-rate volatility: A multivairate latent-factor ARCH model. *Journal of Applied Econometrics* 4: 1–22.

Drost, F.C., and C.A.J. Klaassen. 1997. Efficient estimation in semiparametric GARCH models. *Journal of Econometrics* 81: 193–221.

Drost, F.C., and T.E. Nijman. 1993. Temporal aggregation of GARCH processes. *Econometrica* 61: 909–927.

Engle, R.F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50: 987–1008.

Engle, R.F., and T. Bollerslev. 1986. Modeling the persistence of conditional variances. *Econometric Reviews* 5: 1–50.

Engle, R.F., and G. González-Rivera. 1991. Semiparametric ARCH models. *Journal of Business and Economic Statistics* 9: 345–359.

Engle, R.F., D.M. Lilien, and R.P. Robins. 1987. Estimating time varying risk premia in the term structure: The ARCH-M model. *Econometrica* 19: 3–29.

Engle, R.F., and V.K. Ng. 1993. Measuring and testing the impact of news on volatility. *Journal of Finance* 48: 1749–1778.

Engle, R.F., V.K. Ng, and M. Rothschild. 1990. Asset pricing with a FACTOR-ARCH covariance structure: Empirical estimates for treasury bills. *Journal of Econometrics* 45: 213–237.

Engle, R.F., and K. Sheppard. 2001. Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH. Working Paper No. 8554. Cambridge, MA: NBER.

Fama, E.F. 1965. The behavior of stock market prices. *Journal of Business* 38: 34–105.

Glosten, L.R., R. Jagannathan, and D.E. Runkle. 1993. On the relation between the expected value and the volatility of the nominal excess returns on stocks. *Journal of Finance* 48: 1779–1801.

Hafner, C. 1998. *Nonlinear time series analysis with applications to foreign exchange rate volatility.* Heidelberg: Physica.

Hall, P., and Q. Yao. 2003. Inference in ARCH and GARCH models with heavy tailed errors. *Econometrica* 71: 285–317.

Hansen, B.A. 1991. GARCH(1,1) processes are near epoch dependent. *Economics Letters* 36: 181–186.

Hansen, P.R., and A. Lunde. 2005. A forecast comparison of volatility models: Does anything beat a GARCH (1,1). *Journal of Applied Econometrics* 20: 873–889.

Härdle, W., and A.B.. Tsybakov. 1997. Locally polynomial estimators of the volatility function. *Journal of Econometrics* 81: 223–242.

Härdle, W., A.B.. Tsybakov, and L. Yang. 1996. Nonparametric vector autoregression. Discussion Paper, SFB 373. Berlin: Humbodt-Universität.

Jensen, S.T., and A. Rahbek. 2004. Asymptotic normality of the QMLE of ARCH in the nonstationary case. *Econometrica* 72: 641–646.

Kim, W., and O. Linton. 2004. A local instrumental variable estimation method for generalized additive volatility models. *Econometric Theory* 20: 1094–1139.

Lee, S.-W., and B.E. Hansen. 1994. Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory* 10: 29–52.

Linton, O.B. 1993. Adaptive estimation in ARCH models. *Econometric Theory* 9: 539–569.

Linton, O.B., and E. Mammen. 2005. Estimating semiparametric ARCH(∞) models by kernel smoothing methods. *Econometrica* 73: 771–836.

Lumsdaine, R. 1995. Finite-sample properties of the maximum likelihood estimator in GARCH(1,1) and IGARCH(1,1) models: A Monte Carlo investigation. *Journal of Business and Economic Statistics* 13: 1–10.

Lumsdaine, R.L. 1996. Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models. *Econometrica* 64: 575–596.

Mandelbrot, B. 1963. The variation of certain speculative prices. *Journal of Business* 36: 394–419.

Masry, E., and D. Tjøstheim. 1995. Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence and asymptotic normality. *Econometric Theory* 11: 258–289.

Meitz, M., and P. Saikkonen. 2004. Ergodicity, mixing, and existence of moments of a class of Markov models with applications to GARCH and ACD models. Working Paper Series in Economics and Finance No. 573. Stockholm School of Economics.

Merton, R.C. 1973. An intertemporal capital asset pricing model. *Econometrica* 41: 867–887.

Mikosch, T., and C. Starica. 2000. Limit theory for the sample autocorrelations and extremes of a GARCH (1,1) process. *Annals of Statistics* 28: 1427–1451.

Nelson, D.B. 1990. Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory* 6: 318–334.

Nelson, D.B. 1991. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* 59: 347–370.

Nelson, D.B., and C.Q. Cao. 1992. Inequality constraints in the univariate GARCH model. *Journal of Business and Economic Statistics* 10: 229–235.

Newey, W.K., and D.G. Steigerwald. 1997. Asymptotic bias for quasi-maximum-likelihood estimators in conditional heteroskedasticity models. *Econometrica* 65: 587–599.

Pagan, A.R., and G.W. Schwert. 1990. Alternative models for conditional stock volatility. *Journal of Econometrics* 45: 267–290.

Pagan, A.R., and Y.S. Hong. 1991. Nonparametric estimation and the risk premium. In *Nonparametric and semiparametric methods in econometrics and statistics*, ed. W. Barnett, J. Powell, and G.E. Tauchen. Cambridge: Cambridge University Press.

Robinson, P.M. 1991. Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *Journal of Econometrics* 47: 67–84.

Robinson, P.M., and P. Zaffaroni. 2006. Pseudo-maximum likelihood estimation of ARCH(∞) models. *Annals of Statistics* 34: 1049–1074.

Sentana, E. 1998. The relation between conditionally heteroskedastic factor models and factor GARCH models. *Econometrics Journal* 1: 1–9.

Yang, L., W. Härdle, and J.P. Nielsen. 1999. Nonparametric autoregression with multiplicative volatility and additive mean. *Journal of Time Series Analysis* 20: 579–604.

Arima Models

A. C. Harvey

Autoregressive integrated moving-average (ARIMA) models are models which can be fitted to a single time series and used to make predictions of future observations. They owe their popularity primarily to the work of Box and Jenkins (1970), who defined the class of ARIMA and seasonal ARIMA models and provided a methodology for selecting a suitable model from that class.

The ARIMA class of models emerged as the result of a synthesis between the theory of stationary stochastic processes and certain ad hoc forecasting procedures based on the discounting of past observations. From the theoretical point of view the ability of an autoregressive-moving average (ARMA) process to approximate any linear stationary process was well known. On the other hand, it had been shown by Muth (1960) that the forecasts generated by the exponentially weighted moving average (EWMA) procedure, i.e.

$$\hat{y}_{t+1/t} = \lambda y_t + (1-\lambda)\hat{y}_{t/t-1}, \tag{1}$$

where $\hat{y}_{t+1/t}$ is the prediction of y_{t+1} made at time t and λ is the smoothing constant, are identical to the optimal one step ahead forecasts which result when the differenced observations are modelled by a first order moving average process, i.e.

$$\Delta y_t = \xi_t + \theta_1 \xi_{t-1}, \tag{2}$$

where ξ_t is a random disturbance term, Δ is the first difference operator and the MA parameter, θ, is equal to $\lambda - 1$. This result was extended to show that the forecasts produced by Holt's local linear trend procedure are the same as those given by a model in which second differences follow a second-order moving average process,

$$\Delta^2 y_t = \xi_t + \theta_1 \xi_{t-1} + \theta_2 \xi_{t-2}; \tag{3}$$

see Theil and Wage (1964), Nerlove and Wage (1964), and Harrison (1967). The nature of the synthesis effected by Box and Jenkins was to formulate a class of models in which the dth difference of the observations was taken to be stationary and hence capable of approximation by an ARMA process with p autoregressive parameters, $\phi_1, \ldots, \phi_p$ and q moving average parameters $\theta_1, \ldots, \theta_q$, i.e.

$$\Delta^d y_t = \phi_t \Delta^d y_{t-1} + \cdots + \phi_p \Delta^d y_{t-p} + \xi_t + \theta_1 \xi_{t-1} + \cdots + \theta_q \xi_{t-q}. \tag{4}$$

The specification of Eq. 4 is denoted by writing it as ARIMA (p, d, q). Thus Eq. 2 is ARIMA (0, 1, 1) while Eq. 3 is ARIMA (0, 2, 2).

Given the ARIMA class of models, it was necessary to provide a methodology for choosing a suitable model from the class. Box and Jenkins (1970) proposed a model selection cycle based on three stages: identification, estimation and diagnostic checking. In the identification stage tentative choices are made for the values of p, d and q using statistical tools such as the correlogram and the sample partial autocorrelation function. Given a specification of these values, the parameters in the model are estimated by maximum

likelihood (ML) or an approximation to maximum likelihood. The residuals from the model are then subject to diagnostic checking to determine if they appear to be approximately random. If the model fails these diagnostic checks the complete cycle is repeated, starting with an attempt to identify a new model. Once a suitable model has been fitted, it can be used to make predictions of future observations, together with estimates of the corresponding mean square errors.

ARIMA models of the form (Eq. 4) are not, in general, appropriate for modelling monthly and quarterly observations as these typically contain a seasonal pattern. However, Box and Jenkins (1970, ch. 9) observed that taking an EWMA of the observations combined with an EWMA of the observations on the current month in previous years, not only produced a viable forecasting procedure but could also be nationalized by the stochastic process

$$\Delta\Delta_s y_t = (1 + \theta L)(1 + \Theta L^s)\xi_t \qquad (5)$$

where Δs is the seasonal difference operator, and θ and Θ are parameters. Generalizing (Eq. 5) gives the class of multiplicative seasonal ARIMA processes, in which a model of order $(p, d, q) \times (P, D, Q)_s$ is specified as

$$\phi(L)\Phi(L^s)\Delta^d\Delta_s^D y_t = \theta(L)\Theta(L^s)\xi_t, \qquad (6)$$

Where $\varphi(L)$, $\Phi(L^s)$, $\theta(L)$ and $\Theta(L^s)$ are polynomials in the lag operator of order p, P, q and Q respectively. The methodology for selecting a model for the seasonal ARIMA class is essentially the same as that developed for the ARIMA class.

The application of the model selection methodology advocated by Box and Jenkins (1970) is not without its problems. Unless the sample size is very large, which it rarely is in economics, it is difficult to identify an ARIMA model of any degree of complexity using the correlogram and the sample partial autocorrelation function. These difficulties become even more acute when the observations have been differenced. One way of avoiding these problems is to select models by an automatic procedure, using a measure of goodness of fit such as the Akaike Information Criterion (AIC). This approach is now quite common, although it does move away from the spirit of the work of Box and Jenkins (1970), which emphasized the need for judgement on the part of the statistician.

A more radical criticism of Box–Jenkins methodology concerns the suitability of the ARIMA class itself. There is no overwhelming reason why an economic time series should, after an appropriate amount of differencing, be stationary. Furthermore, even if the stationarity assumption is a reasonable one for a differenced series, it does not follow that approximating the differenced series by an ARMA (p, q) process will necessarily lead to a model with desirable properties for forecasting. Some illustrations of this point can be found in Harvey and Todd (1983) and Harvey (1985). Thus while the ARIMA class may often be too restrictive because of its reliance on stationarity, it can also be argued that it is too general. Given the difficulties which arise in applying the Box-Jenkins methodology, it follows that there is ample scope for selecting an inappropriate model. As the examples cited by Jenkins (1982) show, the use of an automatic model selection procedure is only likely to make matters worse.

Recent work has suggested an alternative to ARIMA models, based on the idea that the components known to exist in economic time series, for example trends, seasonals and perhaps even cycles, are modelled explicitly. These components are unobserved but may be handled statistically by means of the state space form as in, say, Kitagawa (1981) and Harvey and Todd (1983). Thus more a priori information is put into the initial specification and the model selection methodology is closer to that of econometrics; see Harvey (1985). Following the terminology of stimultaneous equation systems in econometrics, Engle (1978) has termed such models 'structural' models. If the model is linear, the 'reduced form' is an ARIMA process. Within this framework the reduced form provides a valid means of constructing forecasts, but it does not provide any direct information which can be used to describe the nature of the series in terms of components of interest.

See Also

- ▶ Autoregressive and Moving-Average Time-Series Processes
- ▶ Stationary Time Series
- ▶ Time Series Analysis

References

Box, G.E.P., and G.M. Jenkins. 1970. *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Engle, R.F. 1978. Estimating structural models of seasonality. In *Seasonal analysis of economic time series*, ed. A. Zellner, 281–308. Washington, DC: Bureau of the Census.

Harrison, P.J. 1967. Exponential smoothing and short-term sales forecasting. *Management Science* 13: 821–842.

Harvey, A.C. 1985. Trends and cycles in macroeconomic time series. *Journal of Business and Economic Statistics* 3: 216–227.

Harvey, A.C., and P.H.J. Todd. 1983. Forecasting economic time series with structural and Box–Jenkins models: A case study (with discussion). *Journal of Business and Economic Statistics* 1: 229–315.

Jenkins, G.M. 1982. Some practical aspects of forecasting in organisations. *Journal of Forecasting* 1: 3–21.

Kitagawa, G. 1981. A nonstationary time series model and its fitting by a recursive filter. *Journal of Time Series Analysis* 2: 103–116.

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

Nerlove, M., and S. Wage. 1964. On the optimality of adaptive forecasting. *Management Science* 10: 207–224.

Theil, H., and S. Wage. 1964. Some observations on adaptive forecasting. *Management Science* 10: 198–206.

Aristotle (384–322 BC)

M. I. Finley

Aristotle (born Stagira, 384 BC, died Chalcis, 322), spent twenty years from the age of seventeen at Plato's Academy in Athens, to which city he returned in 335 to establish his own school, the Lyceum. He presided over the Lyceum until the death of Alexander the Great (whom he had once tutored) in 323. He then left Athens and died shortly thereafter.

Aristotle has been rightly called a 'universal genius': his works range over formal logic, epistemology and metaphysics; among the natural sciences, physics, meteorology and zoology; also ethics, politics, rhetoric and aesthetics. In each his contribution was a major one, as he defined and codified the subject-matter and indeed created much of the language required for scientific and philosophical discourse. He also established the first wide-ranging research organization, which collected masses of data that Aristotle and his associates employed for their systematic analyses.

The subject notably absent from this great corpus of research and publication is economics. The pseudo-Aristotelian work called *Oikonomikos* is no exception. Apart from the fact that it is a relatively late concoction, almost certainly not a single work in origin (whenever that was), the title and Book 3 (known only from a medieval Latin version) represent a type of literature on 'household management', now best represented from classical antiquity by the *Oikonomikos* of Xenophon, written in the first half of the 4th century BC, in which 'management' of the mistress and slaves of the household occupy a central role, but what we call 'economics' none at all. The other two books of the pseudo-Aristotelian compilation deal anecdotally with public revenue, but in large part only with the devices, based on force and fraud, employed by tyrants and other rulers in order to squeeze funds out of their subjects. In a rudimentary sense, therefore, there is an economic component in the work, but neither analysis nor any general conceptions.

For the latter there are only two relevant passages in the Aristotelian corpus, both of them digressions. One is in the *Politics* (1256a1–58b8) in the context of the 'natural' and 'unnatural' modes of acquiring wealth, the other in the *Nicomachean Ethics* (1132b20–34a24) in the context of the forms of justice. There has been a serious, though intermittent, modern discussion of these passages, chiefly among historians of economic thought, but there was no visible interest in antiquity. The period of paramount practical interest in Aristotle's economics was the later Middle

Ages, from the early 13th century on, with Thomas Aquinas as the leading spirit. That was the time when Aristotle was both the great authority for the Church's assault on usury, for which the textual basis was firm and indeed obvious, and the authority for the doctrine of 'just price', which was in fact not Aristotelian but the consequence of a mistranslation (or at least a misinterpretation) by his Latin translators.

The digression in the *Politics* begins by establishing five means of 'natural' acquisition – pasturage, agriculture, hunting, fishing and, surprisingly to us, piracy; proceeds to indicate that as human groups became larger it became necessary to import necessaries lacking locally; argues that money was then invented to facilitate such acquisitions, then that money became converted into a good in itself and that its acquisition through profit, called *chrematistics*, was unnatural, with the taking of interest the worst of all. There is no concern here with how value in exchange is determined. For that one turns to the *Ethics*, where, after distinguishing distributive from corrective justice, Aristotle proceeds to digress about justice in exchange. His problem is the achievement of justice in the determination of exchange values, and the few pages are repetitive and unclear, as if the author were thinking aloud in a discussion or lecture. In consequence, virtually every translator and commentator since the Middle Ages has 'interpreted' Aristotle's thought to fit his own notions. The key sentences are these: 'There will therefore be genuine reciprocity when (the products) have been equalized, so that as farmer is to shoemaker, so is that of the shoemaker's product to that of the farmer's.' In that way, there will be no excess, which would be immoral, but 'each will have his own' (1133a33–b3).

This is repeated within a few lines and there can be no question that Aristotle meant 'as farmer is to shoemaker' to be taken literally. But to do so is intolerable under conventional economic thinking. Most commentators have therefore transmuted the thinking, and in the process they have reduced Aristotle's economic ideas to insignificance. No wonder that Schumpeter (1954, p. 57) dismissed Aristotle's analysis as 'decorous, pedestrian, slightly mediocre, and more than slightly pompous common sense'. However, on a straight reading of Aristotle's words, the conclusion seems clear to me that he never pretended to examine the price mechanism or any other aspect of market exchange *as it was practised*. He was offering a normative ethical analysis: much that went on in practice was unethical on his definition and therefore outside his discourse.

In sum, there is no economic analysis in Aristotle, not even in intention; judgements of his performance on that score or attempts to interpret his words so as to rescue them as economic analysis are doomed from the outset. In the more then fifteen years since I published this exposition at some length, I have seen no acceptable refutation of it in neoclassical economic terms, and I believe none to be possible. A more serious effort has been made by some Marxists, most powerfully in a sophisticated polemic by Meikle (1979), who argues that a Marxist view (and only a Marxist view) warrants a positive evaluation of Aristotle's efforts at economic analysis. I remain unpersuaded, firstly because the underlying proposition that Aristotle's age saw the rise for the first time of a genuine system of commodity production, which Aristotle appreciated and sought to grapple with, is one I hold to be historically false; secondly because Meikle fails to consider the critical phrase, 'as farmer is to shoemaker', which I believe undermines his interpretation. And there the debate stands.

See Also

► Chrematistics

Bibliography

Aristotle. *Nicomachean Ethics*. Trans. H. Rackham. London/Cambridge, MA: Heinemann/Harvard University Press (The Loeb Classical Library), revised ed., 1934.

Aristotle. *Politics*. Trans. E. Barker. Oxford: Clarendon, 1946.

Finley, M.I. 1970. Aristotle and economic analysis. *Past and Present* 47: 5–25.

Langholm, O. 1983. *Wealth and money in the Aristotelian tradition*. Oslo: Universitetsforlaget.

Langholm, O. 1984. *The Aristotelian analysis of usury*. Oslo: Universitetsforlaget.

Meikle, S. 1979. Aristotle and the political economy of the polis. *Journal of Hellenic Studies* 99: 57–73.
Schumpeter, J. 1954. *History of economic analysis*. New York: Oxford University Press.

Arms Races

Dagobert L. Brito and Michael D. Intriligator

Abstract

We analyse arms races for an environment in which social, human and intellectual capital are more important than physical capital. The Richardson model can be used to analyse the Anglo–German naval race before the First World War and the US–Soviet missile race during the Cold War; in both cases the economic constraint associated with acquiring weapons was the binding constraint. Previously, human and social capital were more important components of military power. Modern technology has reduced the importance of the economic constraints associated with acquiring physical capital. Our model of such a process suggests that a stable equilibrium is unlikely.

Keywords

Arms races; Arms trade; Cold War; Human capital; Increasing returns to scale; Lotteries; Public goods; Returns to scale; Richardson model of arms races; Risk; Slavery; Social capital; Technical change; Terrorism; Uncertainty

JEL Classifications

N4

The traditional literature on arms races starts with the Richardson model (named after Lewis Fry Richardson, 1881–1953, British polymath who made fundamental contributions to the mathematical analysis of war, to weather forecasting, and to measuring the length of coastlines and borders). The Richardson model is a descriptive model of the dynamic processes of interaction in an arms race. The model is summarized by two differential equations describing the rate of change over time of weapon stocks in each of two countries, 1 and 2. Let $w_1(t)$ represent the stock of weapons for country 1 and $w_2(t)$ represent the stock of weapons for country 2 at time t. In the Richardson model the rate of change of weapon stocks at time t is given by

$$\begin{aligned} \dot{w}_1(t) &= a_1 w_2(t) + b_1 w_1(t) + c_1 \\ \& \\ \dot{w}_2(t) &= a_2 w_1(t) + b_2 w_2(t) + c_2 \end{aligned} \tag{1}$$

According to these coupled differential equations, the accumulation of weapons in country 1 can be described as the sum of three separate influences. First is the 'defence term', a_1, where the accumulation of weapons is influenced positively by the stock of weapons of the opponent, w_2(t), representing the need to defend oneself against the opponent. Second is the 'fatigue term', b_1, where the accumulation of weapons is influenced negatively by one's own stock of weapons, representing the economic and administrative burden of conducting the arms race. Third is the 'grievance term', c_1, representing all other factors influencing the arms race, whether historical, institutional, cultural, or derived from some other source. The dynamics of the arms accumulation equation for country 2 are symmetrical.

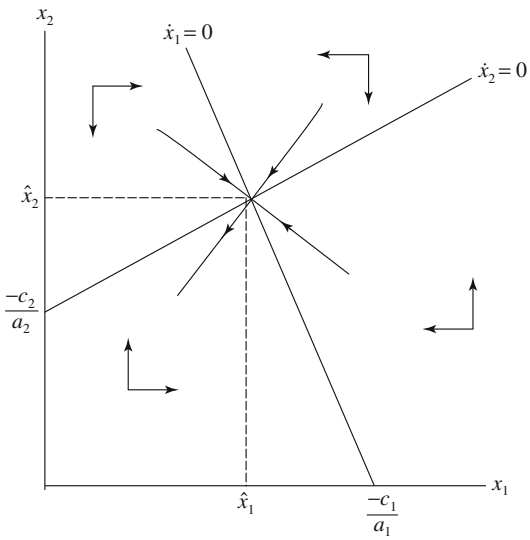
During the Cold War, Richardson's equations attracted much interest among political scientists, economists and others interested in the arms race. One of the questions of interest was the stability of the arms race. There are three schools of thought about the stability of armament races. One is that armaments races have a stable equilibrium. A second belief is that armaments races are unstable, a belief often seen in the popular press, which holds that unless some agreement is reached weapon stocks will increase in an ever-accelerating spiral that must ultimately lead to bankruptcy or nuclear holocaust. A third view is that a stable equilibrium may exist, but that the

stability may only be a local property, so that a large disturbance of the system, such as the introduction of a new weapons system, which may set off an armaments race, either positive (leading to larger and larger weapons stocks) or negative (leading to major decreases in weapons).

The first two questions could be addressed by using the parameters of the model to calculate the roots and check for stability. The third question requires that the underlying process that led to these differential equations be modelled. Much of the theoretical work on the arms race is in the Richardson tradition of explaining the arms race was attempt to estimate these parameters empirically or to find theoretical reasons for constraining the magnitudes (as discussed in the Intriligator 1982, survey paper). The third question was addressed by research that derived the dynamics of arms accumulation in a model based on the axioms of rational choice, on the assumption that each country can be modelled as a single rational actor. Brito (1972) and Intriligator (1975) each obtained a general set of equations describing an arms race, of which the Richardson model is one special case.

In Fig. 1, $\hat{x}_1$ and $\hat{x}_2$ are the stable equilibrium.

The Richardson paradigm was the central focus of research on arms races. During the Cold War, the build-up of bombers and then missiles by the United States and the Soviet Union was, or should have been, the most important concern as it had the potential of destroying civilization as we know it, if not mankind. This danger was reduced with the end of the Cold War and the later dissolution of the Soviet Union, which ended the US–Soviet arms race. The new environment may not be well characterized by the Richardson paradigm; this article describes the changes and suggests a new approach to the formulation of a model of arms races.



Arms Races, Fig. 1

The Changing Nature of Arms Races

There have been several major changes in the nature of the arms race since the early 1990s. The most important has clearly been the end of the Cold War. This epochal change began with the demise of the Warsaw Pact in 1989 and ended with the dissolution of the Soviet Union in December 1991. The result has been the end of the global East–West arms race of the Cold War period, when it dominated global politics. Among the implications of this profound change have been drastic reductions in arms expenditures by the member states of the former Soviet Union and its former allies, accompanied by relatively smaller reductions in arms expenditures by the United States prior to the Afghanistan and Iraq wars and by its allies in NATO. As a result, the United States is currently by far the world leader in expenditures on arms, spending almost as much as the rest of the world combined.

Another major change since the mid-1990s has been the substantial increases in arms expenditures by China and its neighbouring states in east and south-east Asia. In China, the reforms that started as a result of Deng Xiaoping's four modernizations of 1978 profoundly changed the course of the country and its economy and society. The last of these four modernizations was that of the military, which led to the rapid modernization of the Chinese People's Liberation Army (PLA), involving the deployment of newer weapons and major expenditures on arms. The neighbouring nations of east and south-east Asia have reacted

to these developments in China by increasing their own arms expenditures. As a result, this region is witnessing major increases in arms, including substantial arms imports.

The India–Pakistan arms race also continues with both qualitative and quantitative arms developments, both nations having demonstrated their nuclear weapons capabilities in tests conducted in May 1998. In both cases, third parties have played an important role. China has shared nuclear and missile technology with Pakistan, and Pakistan, in turn, has been a major actor in the proliferation to nuclear technology to North Korea and Iran.

In the Middle East, the United States has provided Saudi Arabia with weapons, given financial and military assistance to both Israel and Egypt, and has shared antimissile defence technology with Israel. While Russia can no longer afford to support the former client states of Soviet Union, it appears to be willing to sell weapons technology to any country that can afford it for purely commercial, as opposed to diplomatic or military, purposes.

An important change of recent years has been the appearance of certain newer or evolving regional arms races or arms build-ups. One is the important arms race is that involving the nations of the Gulf, including Iraq, Iran, Syria, Saudi Arabia, Kuwait and the Gulf States, that both was stimulated by and resulted in wars in the region, including the Iran–Iraq war and the Iraqi invasion and annexation Kuwait, resulting in a war to liberate it and the subsequent US-led invasion and occupation of Iraq. The major suppliers of weapons to all parties in the region except Iran are the United States and its European allies. Second, there have also been arms build-ups among the states of the former Soviet Union that are seeking to preserve their independence through their military capabilities. A third type of arms build-up is that in the former Warsaw Pact states of central and eastern Europe that have joined NATO, or hope to do so, and that have to upgrade their weapons capabilities to become members of the alliance.

The major weapons states have played an important role in fuelling these and other regional arms races through arms exports, including the disposal of surplus weapons in the post-cold war period. The United States, Russia, Germany, Britain and France are the leading suppliers of surplus weapons, while Turkey, Greece, Pakistan, Morocco and a number of Middle East countries are the main recipients of such weapons.

Impacts of Recent Changes on Stability

These changes in arms races since the mid-1990s have had important impacts on the stability of both the regional and global systems. As a result of these changes, we believe that there are probably greater instabilities today than those of the earlier Cold War period.

Consider first the principal antagonists of the Cold War. Where there had earlier been two ‘superpowers’, now there is only one as measured by arms expenditures and military capabilities, namely, the United States. Russia has assumed most of the Soviet weapons of mass destruction and the associated responsibilities involved with such weapons. The continued presence of nuclear weapons in Russia and the United States, albeit at lower levels, is probably adequate for mutual deterrence, but there are great dangers inherent in the current unstable political, economic, and social situation in Russia. The result could be a loss of effective control of weapons of mass destruction, with the possibility of an accidental or inadvertent launch of such weapons. The disquieting similarities between Russia today and Germany in the Weimar Republic period between the wars, including loss of empire, inflation, depression and the destruction of the middle class, suggest the possibility of the emergence of a new authoritarian leader in Russia, which would create additional instabilities.

Another major threat to stability at both global and regional levels is the proliferation of weapons of mass destruction. There is now much greater worldwide access to technology and the required material for nuclear, chemical, and biological weapons stemming, in part, from the collapse of the Soviet Union and the desperate situation of its military and scientific establishment. There are also the chains of proliferation that started with

the United States and continued with the Soviet Union, the United Kingdom, France, China, India, and Pakistan, and that could continue to other nations, including Iran and other nations of the Gulf region.

Yet another threat to stability in the post-Cold War world is that of terrorists using various weapons of mass destruction. Sub-national groups, motivated by extreme ideologies, religious fanaticism, or other causes, have much greater access to such weapons on world markets. Large urban centres and freedoms of speech, travel, assembly, and the press have made modern societies highly vulnerable to possible terrorist attack. This was clearly demonstrated on September 11, 2001.

Beyond the Richardson Paradigm

Until the East–West arms race of the Cold War period, most arms races were naval. Until the 20th century, armies were highly labour-intensive institutions with relatively little capital. Roman soldiers furnished their own equipment until the late Republic. Feudal armies also furnished their own equipment, where the obligation of a fief holder under military tenure was to furnish a certain number of knights and men at arms for a given number of days a year and to provide arms and horses for these men. The key element in deploying military power at that time was the organization of the state and its ability to raise revenue. The possibility of organizing and disciplining free men to serve as heavy infantry was the key to the Greek and Republican Roman armies. Heavy infantry required a body of free men willing to serve. It is very difficult to find examples of heavy infantry manned by professional soldiers except in circumstance where the state had the ability to tax effectively, such as the early Roman Empire and European states after the 16th century.

In hindsight, however, the Richardson paradigm of competitive accumulation of weapons, though important, was limited. The Anglo–German naval race that first attracted Richardson's attention played a very minor role in the First World War. After the indecisive battle of Jutland in 1916, both battle fleets were inactive and the important naval element was the German use of U-boats.

The other important arms race of the 20th century that fits the Richardson paradigm was the nuclear arms race between the United States and the Soviet Union. Fortunately, because of mutual assured destruction, these weapons were never used and the downfall of the Soviet Union was largely the result of the failure of its institutions.

Arms races did not play a major role in the Second World War. British aircraft manufacturers increased the stock of fighter planes during the Battle of Britain. The United States did not fully gear up for a war economy until after Pearl Harbor, and Soviet war production came from factories they moved east of the Urals. Even German production was increasing until the very end of the war.

In recent years technological change has also called into question the Richardson paradigm. Constant or increasing returns to scale have always created difficulties for economic theory. An economy with constant returns to scale is indeterminate with respect to the scale size of firms, and it is necessary to appeal to some fixed factor to determine the size of the economy. Increasing returns to scale leads to monopolies constrained only by demand. Firm behaviour then becomes strategic and none of the standard welfare theorems that hold in competitive markets apply. Thus, it is not surprising that increasing returns to scale in an arms race can lead to very different results than constant or decreasing returns to scale.

Increasing returns to scale in the technology of arms production is more likely to occur with newer types of 'smart' weapons that rely heavily on electronics, computers, software, and so forth. In producing weapons with such a large informational component, it is likely that increasing the scale of the production process will make production more efficient. Nations producing arms may sell weapons even when these sales may be contrary to their foreign policy. The drive to lower weapons unit costs through greater sales gives

momentum to foreign arms sales that can even conflict with diplomatic or political goals. An example may be the decision of the United States to lift its embargo on arms sales to Latin America at the urging of weapons producers.

Another consequence of technological change is that new technologies have made nuclear weapons and missiles feasible for most nation states, and some of these technologies have valid non-military applications. North Korea with an annual GDP of US$40 billion has acquired nuclear weapons and is ready to test the Taepondong-2, a missile that can reach the United States or, as the North Koreans claim, put a satellite in orbit. Iran is developing the capability of enriching uranium, a capability that can be used to produce fuel or bombs. As of 2006, the developed world is trying to prevent the test of the missile by North Korea and the acquisition of the capability of enriching uranium by Iran. Technological change has forced the developed world into the position of trying to deny countries in the developing world technologies that the developed world possesses and that have plausible non-military use.

As discussed above, social capital has been a very important element in the ability of a state to mobilize its resources and project power. Social capital includes not only the tangible institutions that the state has to tax, to conscript and to mobilize resources, but also less tangible institutions such as the relations of the members of the state to each other and to the state. States with sharp class, ethnic or caste distinctions may find it difficult to mobilize effectively to project force. During the American Civil War, the institution of slavery kept the South from mobilizing the members of its population that were black, and gave President Lincoln the political advantage of defining the war to be against the institution of slavery. In present day Iraq, ethnic differences have made it very difficult to organize an Iraqi national army.

Among the components of social capital are the common values of the society and its institutions. One important element of social capital familiar to most economists, but largely neglected in the arms race literature, is the attitude of the society towards risk and uncertainty. One very important question is how a society views a lottery that will cost a specific member of society his or her life with certainty to be equivalent to a lottery in which 1,000 individuals face one chance in a thousand of dying. It has long been noted by scholars in such fields as public finance, law, and economics that people in the United States are willing to spend more resources to save a specific individual than an individual who is a statistical abstraction. This element of social capital is reflected in how the United States conducts war, but it is not shared by other cultures.

There is widespread use of suicide bombers in current conflicts in Palestine and Iraq. Although this is a new phenomenon in recent history, most of the elements are not new. In the Second World War Japan sent young pilots on kamikaze suicide missions while the United States was willing to send bomber crews over Germany knowing that few would survive and there would be civilian casualties. The probability that a bomber crew would survive a full tour of duty was small. There may be some substantive difference between the Palestinians being willing to send a young man to kill himself to induce terror among the Israelis and the Doolittle raid where 16 bombers attacked the Japanese home islands in 1942 for psychological purposes; but it seems that the difference is that, whereas Western cultures are willing to sacrifice individuals for the common good as long as the sacrifice is a lottery, some other cultures are willing to sacrifice specific individuals. This difference changes the war-making potential of the different cultures.

To illustrate with another example, the Japanese supply of trained pilots was seriously depleted during the battle of Midway in 1942 and subsequent naval engagements. The Japanese were not able to compete with the Americans in training new pilots. By the Marianas campaign the Japanese were no match for the Americans, and the Japanese resorted to using untrained pilots as kamikazes to attack the American fleet. This example illustrates the role of various forms of social capital in war. The more open and egalitarian American society allowed the United States to train pilots as it had a larger pool to draw from than the more structured and hierarchical Japanese society. However, the

advantage of this type of American social capital was offset in part by the fact that Japanese society was willing to sacrifice specific individuals. American pilots were better trained and had more human capital; the willingness of Japanese society to sacrifice specific Japanese pilots was a different form of social capital.

Richardson's world was one in which dreadnoughts and battlecruisers would steam into battle planned by admirals who had studied Admiral Thayer Mahan (1840–1914, US naval officer and geostrategist who was influential on the US building a modern naval fleet, acquiring overseas naval bases, and building the Panama Canal) and other theorists. The US–Soviet arms race was also a very intellectual process that was based on very sophisticated doctrines and involved weapons systems that were highly quantifiable. The conflicts we now face, by contrast, are very different. They involve state and non-state entities, and the means of deploying force are highly asymmetrical. Fighter planes carrying GPS guided bombs are used against terrorists who employ suicide bombers and can use the internet to transmit pictures of the decapitation of prisoners. Modelling such phenomena is the task for the next generation. What we propose to do is offer a conjecture as to the nature of such processes.

A Conjecture on Arms Race Theory

Assume that the war-making potential of the i-th country can be described by a vector of physical, human, intellectual and social capital, $\mathbf{k_i}$ and a vector of strategies $\mathbf{V}_\mathrm{i}$. Its war-making potential, $x_{i,}$ is given by

$$x_i = \max_{v_i} \ \theta \ (\mathbf{k}_i, \mathbf{v}_i) \tag{2}$$

where the cost of the strategies and other tradeoffs is reflected in the social capital. We conjecture that the intertemporal optimization results in a differential equation of the form

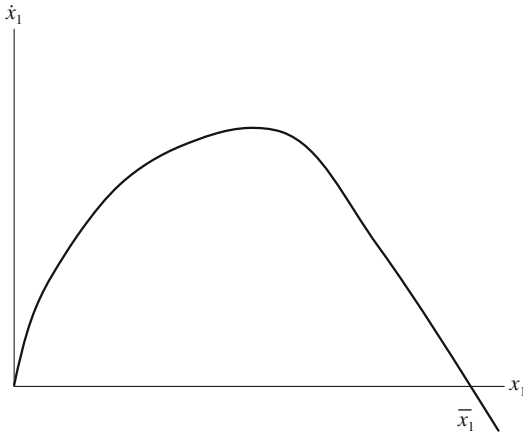
$$\begin{aligned} &\dot{x}_1 = a_1(x_1) + b_1(x_2) + c_1(x_1, x_2) \\ &\& \\ &\dot{x}_2 = a_2(x_2) + b_2(x_1) + c_2(x_1, x_2) \end{aligned} \tag{3}$$

The first term $a_i(x_i)$ reflects the role of the i-th country's war-making potential on the rate of growth of i-th country's war-making potential. In the Richardson model the derivative of this term is negative as it represents the fatigue term. In this model it could well be positive as many of the components of the war-making potential– social, intellectual and human capital –are productive. The second term $b_i(x_j)$ reflects the role of the j-th country's war-making potential on the rate of growth of i-th country's war-making potential. This is analogous to the defence term in the Richardson model. As in the Richardson model, this term is positive. In this model such an assumption is made for two reasons. First, as in the Richardson model, an increase in the war-making potential of the j-th country will be viewed as a threat. Second, and perhaps more important, some of the inputs in the production of x_j, particularly intellectual capital and social capital, are public goods and can be transferred to the competing country. Meiji Japan acquired from the West the technology to build warships and organize a modern navy, and at the present time the technology the North Koreans are using to build nuclear bombs can be traced from the United States through various intermediaries to China, to Pakistan and then to North Korea. The problem of technological transfer is more difficult to control when it is dual use, that is, could be used for civilian as well as military purposes. After all, the Taepondong-2 *could* be used to launch weather satellites. The term, $c_i(x_i, x_j)$ is different from the grievance term in the Richardson model in that it represents the competition of the parties for resources or perhaps even ecological space, and is assumed to be quadratic in order. The derivative is assumed to be positive. If we consider the equation

$$\dot{x}_1 = a_1(x_1) + b_1(0) + c_1(x_1, 0) \tag{4}$$

and if $a_1(0) = 0$ and $b_1(0) = 0$, we would assume that Eq. (4) would behave in a way similar to a biological population growth equation (Fig. 2).

$\bar{x}_1$ is the maximum potential size of Country 1 in the absence of competition. A linear approximation of Eq. (3) is given by



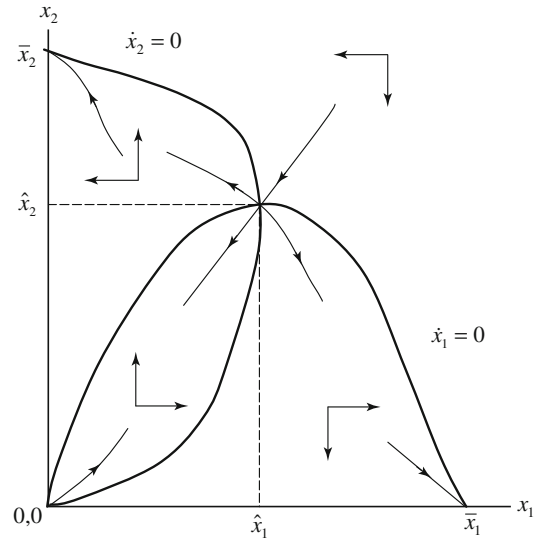
Arms Races, Fig. 2

$$\begin{aligned} \dot{x}_1 &= a_1 x_1 + b_1 x_2 + c_1 (x_1 + x_2)^2 \\ \& \\ \dot{x}_2 &= a_2 x_2 + b_2 x_1 + c_2 (x_1 + x_2)^2 \end{aligned} \tag{5}$$

This is similar to the Richardson equation except for the quadratic term of the common resource constraint. if we assume that a_i, the 'fatigue term', is negative, b_i, the 'defence term', is positive and c_i, the 'resource term', is negative,then we can represent the dynamics of this nonlinear system in the phase diagram in Fig. 3.

Although on the surface this appears to be very similar to the Richardson equation, the variable x_i is war-making potential that is the result of a prior optimization. One of the elements of the prior optimization is social capital, which includes among its elements moral values.

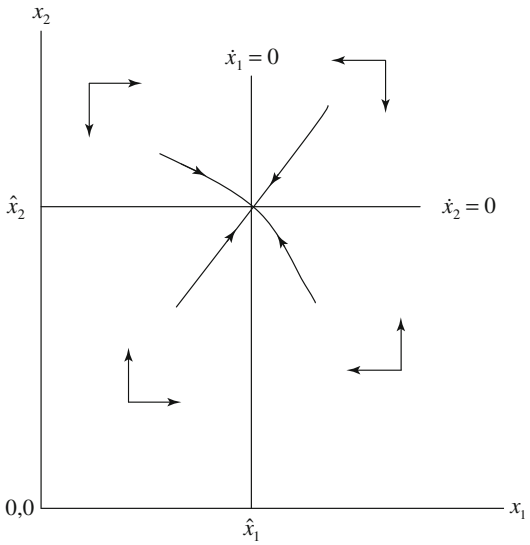
The differential equation system has four equilbria, of which two are stable and two are unstable. The two that are stable $(\overline{x}_1, 0)$ and $(0, \overline{x}_2)$, involve the elimination of one of the parties. Whether this is good or bad depends on the process of the optimizations underlying the dynamical system. Recall that one of the important components of the process is social capital. One realization could be that the social capital of the competing parties would evolve in such a fashion as to eliminate conflict. An example is the transformation of the nation states of Europe, with a thousand-year history of wars, into the European Union. A second, less optimistic, scenario is the complete destruction



Arms Races, Fig. 3

of the weaker party. Again, the crucial element is social capital. Initially, the weaker power many threaten the stronger power by using tactics that are not acceptable to the values of the stronger power – for example, the use of suicide bombers. However, civilization has a thin veneer. Historically, if a country feels that its survival or vital interests are at stake, it will quickly shed its inhibitions. The tactics the British used to suppress the sepoy mutiny were brutal. At Peshawar, 40 sepoys were stood before cannons and blown apart in a public execution. The countries that condemned the German bombing of Guernica in the Spanish Civil War (fewer than 2,000 casualties) firebombed Hamburg (50,000 casualties), Dresden (25,000–35,000 casualties) and Tokyo (100,000 casualties) in the Second World War, and ultimately used atomic weapons on Japanese cities. Before the start of the Gulf War of 1991, US Secretary of State James A. Baker III warned Iraq that the use of weapons of mass destruction by Iraq would result the destruction of Iraq as a modern state.

The third alternative is decoupling. This results in a stable equilibrium (see Fig. 4). The French in Algeria, the United States in Vietnam and the Soviet Union in Afghanistan withdrew because the game was not worth the candle. The partition that appears to be imminent in Palestine, where



Arms Races, Fig. 4

Israel is building a wall to minimize its interaction with the Palestinians, may be an omen of things to come. If interaction with the developing world becomes too costly, the developed world has the alternative of disengaging. Without oil, the Middle East would be no more important than Africa, and conflicts between the Sunnis and Sh'ias would receive the same attention as conflict between the various African tribes. At prices greater than US $45.00 a barrel, technologies exist for the developed world to be self-sufficient in oil. History could repeat itself. An argument can be made that the Muslim world started to decline in the 16th century partly because the opening of alternative trade routes to Asia destroyed the Muslim monopoly on such trade.

Conclusions

The arms race as described by the Richardson paradigm, where nation states arm in a competitive fashion, is a phenomenon that starts with the naval arms race at the end of the 19th century and may have ended along with the US–Soviet arms race after the dissolution of the Soviet Union. Before that time, warfare was not very capitalintensive, and the most important elements in the projection of military power were the human capital of the population and the social capital that enabled countries to mobilize their resources in war.

Richardsonian arms races reflect competition that is constrained by economic resources. Recent developments in technology have broken that link. Technological change has made it possible for a country like North Korea, with an annual GNP of US$40 billion, to acquire nuclear weapons and a missile that may be capable of attacking the United States. The link between economic power and the ability to project military power has been broken. The Richardson paradigm no longer applies. We conjecture the structure of an alternative model. This model suggests three alternatives: cultural convergence, destruction of the weaker party, and decoupling of the conflict. It should be clear that the model is a conjecture based on our intuition, and much work is needed to develop the theoretical foundations of the next arms race paradigm.

See Also

- ▶ Arms Trade
- ▶ Defence Economics
- ▶ War and Economics

Bibliography

Anderton, C. 1985. A selected bibliography of arms race models and related subjects. *Conflict Management and Peace Science* 8: 99–122.

Anderton, C. 1989. Arms race modeling: Problems and prospects. *Journal of Conflict Resolution* 33: 346–367.

Boulding, K. 1978. Future directions in conflict and peace studies. *Journal of Conflict Resolution* 22: 342–354.

Brito, D. 1972. A dynamic model of an armaments race. *International Economic Review* 13: 359–375.

Brito, D., and M. Intriligator. 1995. Arms races and proliferation. In *Handbook of defence economics*, ed. K. Hartley and T. Sandler. Amsterdam: North-Holland.

Brito, D., and M. Intriligator. 1996. Proliferation and the probability of war: A cardinality theorem. *Journal of Conflict Resolution* 40: 204–212.

Brito, D., and M. Intriligator. 1999. Increasing returns to scale and the arms race: The end of the Richardson paradigm? *Defence and Peace Economics* 10: 39–54.

Downs, G., and D. Rocke. 1990. *Tacit bargaining, arms races, and arms control*. Ann Arbor: University of Michigan Press.

Gleditsch, N., and O. Njølstad, eds. 1990. *Arms races, technological and political dynamics*. London: Sage Publications.
Intriligator, M. 1975. Strategic considerations in the Richardson model of arms races. *Journal of Political Economy* 83: 339–353.
Intriligator, M. 1982. Research on conflict theory: Analytic approaches and areas of application. *Journal of Conflict Resolution* 26: 307–327.
Intriligator, M., and D. Brito. 1984. Can arms races lead to the outbreak of war? *Journal of Conflict Resolution* 28: 63–84.
Isard, W. 1988. *Arms races, arms control and conflict analysis*. New York: Cambridge University Press.
Isard, W., and C. Anderton. 1985. Arms race models: A survey and synthesis. *Conflict Management and Peace Science* 8: 27–98.
Nicholson, M. 1999. Review article: Lewis Fry Richardson and the study of the causes of war. *British Journal of Political Science* 29: 541–563.
Richardson, L. 1960. *Arms and insecurity: A mathematical study of the causes and origins of war*. Pittsburgh: Boxwood Press.

Arms Trade

Charles H. Anderton and John R. Carter

Abstract

Arms trade is the transfer of weapons systems, components, technologies, and services across national and territorial borders. Contemporary arms trade occurs in three product categories: major conventional weapons; small arms and light weapons; and weapons of mass destruction. This article briefly surveys the theoretical and empirical arms trade literature. Topics include arms trade data sources, commercial and security motives for weapons exports, competitive and imperfectly competitive models of arms trade, empirical studies of the economic and political effects of arms trade, and arms export controls.

Keywords

Arms trade; Arms exports; Arms export restraints

JEL Classifications

H5

Arms trade is the transfer of weapons systems, components, technologies and services across national and territorial borders. Contemporary arms trade occurs in three product categories: major conventional weapons (MCW), such as fighter aircraft and destroyers; small arms and light weapons (SALW), such as assault rifles, machine guns and improvised explosive devices; and weapons of mass destruction (WMD), such as nuclear, biological and chemical weapons technologies and long-range missile systems. MCW are the dominant form of weapons in interstate wars, while SALW are used intensively by non-state actors in intra-state wars (for example, civil wars) and extra-state conflicts (for example, transnational terrorism). WMD components and technologies proliferate by spreading to states or possibly non-state actors via trade or indigenous production.

Major sources of arms trade data include the US Congressional Research Service for all categories of weapons and arms-related services to developing nations; the Stockholm International Peace Research Institute for MCW; the Norwegian Initiative on Small Arms Transfers and the Graduate Institute of International Studies (Geneva) Small Arms Survey for SALW; and the Monterey Institute's Center for Nonproliferation Studies for WMD proliferation. These sources indicate that, of the world's total arms exports, more than one-half originates in the United States and Russia, and close to two-thirds goes to developing nations (Brauer 2007).

Theories of arms trade have shifted in emphasis over time. Pre-Cold War literature emphasized economic motives, often from a condemnatory 'merchants of death' perspective (see, for example, Engelbrecht and Hanighen 1934). During the Cold War, classic texts focused on domestic and international politics, with some coverage of economic incentives (see, for example, Pierre 1982). Post-Cold War models of arms trade highlight both commercial and security concerns. For example, in Levine and Smith's (1995) model, a

few suppliers export weapons to a large number of price-taking buyers who are involved in dyadic arms rivalries. Suppliers' utility depends on security and producers' profits, while recipients' utility depends on security and consumption. Under certain conditions, commercial gains to arms exporters are offset by security losses because the arms exports create a greater risk of war among recipients. Under other conditions, arms exports reduce war risk, implying both commercial and security gains to suppliers from weapons exports.

Theoretical models of international trade and industrial organization often apply to arms trade (see, for example, Anderton 1996). Competitive models are useful for the study of SALW trade because such weapons are relatively homogeneous and the number of buyers and sellers is large. For MCW and WMD, the number of suppliers is relatively small and products within weapons classes are differentiated. For these weapons, models incorporating economies of scale, technological differences, intermediate products and strategic behaviour are more appropriate.

Some empirical studies investigate the determinants of arms trade (for example, Smith and Tasiran 2005), but most focus on economic and political effects, including the impact on employment, growth and development, arms rivalries, and human rights (see, for example, Grobar et al. 1990; Yakovlev 2005; Sanjian 1999; and Blanton 1999). Perhaps the most important empirical relationship considered is the effect of arms trade on the risk of war. Craft and Smaldone (2003) report that arms imports significantly increase the risk of interstate or intrastate conflict for sub-Saharan African nations. Krause (2004) finds that arms transfers that occur outside of defence pacts increase the risk that recipients will become involved in militarized interstate disputes. Most other studies likewise find that arms exports increase the risk of conflict, but there are exceptions (see Anderton 1995).

Arms exports are typically subject to extensive government influence. Arms trade offsets require an exporting firm to use some of the revenue from arms sales to invest in activities in the importing nation. Brauer and Dunne (2004) report that there is little empirical or case study evidence that arms trade offsets enhance economic development. Some interventions, like subsidies and diplomatic lobbying on behalf of weapons firms, enhance arms exports. Virtually all governments limit arms exports to particular recipients, and various multilateral arms export limitation regimes exist including the Wassenaar Arrangement, the EU Code of Conduct on Arms Exports, the Nuclear Suppliers Group, the Missile Technology Control Regime, and the Australia Group. Brzoska (2004) argues in favour of a multilateral arms export tax in order to reduce arms exports.

Because production and trade are jointly determined economic activities, arms export restraints cannot be understood in isolation from arms production (Brauer 2000). In a competitive market model, reduction of weapons supply through production or export controls can raise the equilibrium world price, creating an incentive for new arms suppliers to enter the market or existing suppliers to circumvent the controls. This suggests that a reduction in weapons demand or an increase in the cost structure of weapons firms is necessary to reduce the number of weapons in the international system in the long run (see, for example, Anderton 1996; Brauer 2000). In Levine and Smith's (1995) imperfect competition model, arms export restraints can benefit suppliers by raising prices and also reduce inefficiencies associated with recipients' arms rivalries. On the assumption that arms sales are taxed, proceeds could be distributed to recipients so that the control regime would Pareto-dominate the outcome with no controls. Such a regime would, however, be vulnerable to cartel-like defections of individual suppliers.

Arms trade involves many direct and indirect economic and political costs and benefits, which suggest a number of broad research themes going forward. First, for the sake of tractability, partial equilibrium analyses of arms trade determinants and effects will continue to dominate the literature. Second, general equilibrium perspectives are beginning to emerge which promise a richer assessment of the nature and effects of arms trade and arms export restraints (see, for example,

Levine et al. 2000). Third, efforts by governments, NGOs, and multilateral organizations to implement Pareto-improving arms trade policies require collective action solutions (Sandler 2000).

See Also

- ▶ Arms Races
- ▶ War and Economics

Bibliography

Anderton, C.H. 1995. Economics of arms trade. In *Handbook of defense economics*, ed. K. Hartley and T. Sandler, Vol. 1. Amsterdam: North-Holland.

Anderton, C.H. 1996. What can international trade theory say about the arms trade? *Peace Economics, Peace Science, and Public Policy* 4: 7–30.

Blanton, S.L. 1999. Instruments of security or tools of repression? Arms imports and human rights conditions in developing countries. *Journal of Peace Research* 36: 233–244.

Brauer, J. 2000. Potential and actual arms production: Implications for the arms trade debate. *Defence and Peace Economics* 11: 461–480.

Brauer, J. 2007. Arms industries, arms trade, and developing countries. In *Handbook of defense economics*, ed. T. Sandler and K. Hartley, Vol. 2. Amsterdam: North-Holland.

Brauer, J., and J.P. Dunne, eds. 2004. *Arms trade and economic development: Theory, policy, and cases in arms trade offsets*. New York: Routledge.

Brzoska, M. 2004. Taxation of the global arms trade? An overview of the issues. *Kyklos* 57: 149–172.

Craft, C., and J.P. Smaldone. 2003. Arms imports in Sub-Saharan Africa: Predicting conflict involvement. *Defence and Peace Economics* 14: 37–49.

Engelbrecht, H.C., and F.C. Hanighen. 1934. *Merchants of death: A study of the international armament industry.* New York: Dodd, Mead.

Grobar, L.M., E.M. Stern, and A.V. Deardorff. 1990. The economic effects of international trade in armaments in the major Western industrialized and developing countries. *Defence Economics* 1: 97–120.

Krause, V. 2004. Hazardous weapons? Effects of arms transfers and defense pacts on militarized disputes, 1950–1995. *International Interactions* 30: 349–371.

Levine, P., S. Sen, and R. Smith, eds. 2000. Special issue: Arms exports, controls and production. *Defence and Peace Economics* 11: 443–548.

Levine, P., and R. Smith. 1995. The arms trade and arms control. *Economic Journal* 105: 471–484.

Pierre, A.J. 1982. *The global politics of arms sales*. Princeton: Princeton University Press.

Sandler, T. 2000. Arms trade, arms control, and security: Collective action issues. *Defence and Peace Economics* 11: 533–548.

Sanjian, G.S. 1999. Promoting stability or instability? Arms transfers and regional rivalries, 1950–91. *International Studies Quarterly* 43: 641–670.

Smith, R.P., and A. Tasiran. 2005. The demand for arms imports. *Journal of Peace Research* 42: 167–181.

Yakovlev, P. 2005. Do arms exports stimulate economic growth? Ph.D. thesis. Department of Economics, West Virginia University.

Armstrong, Wallace Edwin (1892–1980)

C. A. Gregory and James Urray

Born in England in 1892, W.E. Armstrong won an exhibition to Sidney Sussex College, Cambridge before World War I. At the outbreak of war he joined the Royal Medical Corps, but was wounded in action in 1915 and subsequently lost a leg. He returned to Cambridge and completed his degree in the Moral Sciences Tripos in 1918.

In his final year at Cambridge, Armstrong concentrated on psychology and was introduced to W.H.R. Rivers who interested him in anthropology. After completing his degree Armstrong studied anthropology under Haddon and from 1919 carried out field research in Papua New Guinea. Armstrong first worked in South-Eastern Papua, and early in 1921 was engaged by the Papuan government to collect further ethnographic material in this region (Armstrong 1922). He was later appointed Assistant Anthropologist to the government and spent two months on Rossel (Yela) Island in the far east of South-Eastern Papua.

Rossel Island was little known, but had acquired an infamous reputation after a French ship carrying over three hundred Chinese to Australia was wrecked on its coasts in 1858. Nearly all the survivors were killed and eaten by the islanders (Armstrong 1928a, Appendix 1). Anthropologically, the island is of considerable interest as its people speak a non-Austronesian language in contrast to the mostly Austronesian-speaking

inhabitants of the Massim area to the west. Armstrong was probably attracted to the island on account of its ethnological significance, but his initial intention had been to study the islanders' kinship system. In the course of his research he discovered a unique 'monetary' system and concentrated on this aspect of the island's culture. The majority of Armstrong's ethnographic writings are concerned with the monetary system (Armstrong 1923/24, 1924a, b).

In 1922 Armstrong returned to Cambridge, where he was appointed to a temporary lectureship in social anthropology until 1925/26 when, because of changes in the teaching of anthropology, his post was not renewed, and his career in anthropology effectively ended. During the early 1920s Armstrong had become interested in economics and from 1926 to 1939 he acted as supervisor and occasional lecturer in economics at Cambridge. In 1939 he accepted a post as lecturer in economics at the University of Southampton. Eventually, after steady promotion, he became Professor of Economic Theory in the University and retired in 1961. He spent some of his retirement at the University of the West Indies (Armstrong et al. 1974). He died in 1980.

Armstrong's transformation from anthropologist to neoclassical economist did not involve a complete break with his intellectual past. The interest in psychology and *a priori* reasoning which he displayed in his anthropological writings (Urry 1985) equipped him well for his career as a neoclassical theorist. What he did abandon, though, was an interest in empirical research. His contributions to economics were all in the area of pure theory and his arguments illustrated by use of counterfactual examples. For example his book, *Saving and Investment* (1936), explores the logical consequences of the assumption that human beings equate the marginal disutility of labour with the marginal utility of the product. A remarkable feature of this book is the extended use of the 'Robinson Crusoe' model of an imaginary island economy. Actual island economies bear no relation to this imaginary model and Armstrong's anthropological colleagues, such as Malinowski (1921, 1922) and Mauss (1925) were particularly critical of economists for this reason. Armstrong was well aware of these criticisms but never addressed them nor did he concern himself with empirical work or anthropology ever again.

Following the publication of his book (1936) Armstrong turned to the utility controversy. He developed a cardinal theory of utility (1939, 1948) and attempted to dismiss the ordinal theory on logical grounds (1950, p. 119). This involved him in a debate with Little (1950) and Georgescu-Roegen (1954) among others.

While Armstrong's book (1936) has passed largely unnoticed – his Pigovian-inspired theory was, after all, published in the same year as Keynes's *General Theory* – his writings on utility have attracted some attention. Ng (1975), for example, acknowledges his debt to Armstrong.

Armstrong's place in the history of anthropological thought is more secure. His ideas, while now outdated (Liep 1983; Urry 1985), are nevertheless of continued interest. His description of the Rossel Island 'monetary' system was until recently the only primary source on the subject and stimulated much secondary research.

Selected Works

1922. Report on Suau-tawala. *Annual Report 1920–21*. Melbourne: Government Printer.
1923/24. Rossel Island religion. *Anthropos* 18/19: 1–11.
1924a. Rossel Island money: a unique monetary system. *Economic Journal* 34: 423–429.
1924b. Shell money from Rossel Island, Papua. *Man* 24: 161–162.
1928a. *Rossel Island: An ethnological study*. Cambridge: Cambridge University Press.
1928b. Social constructiveness III. *British Journal of Psychology* 18: 366–399.
1936. *Saving and investment*. London: Routledge.
1939. The determinateness of the utility function. *Economic Journal* 49: 453–467.
1948. Uncertainty and the utility function. *Economic Journal* 58: 1–10.
1950. A note in the theory of consumer's behaviour. *Oxford Economic Papers* 2: 119–122.
1951. Utility and the theory of welfare. *Oxford Economic Papers* 3: 259–271.

1953. Marginal preference and the theory of welfare. *Oxford Economic Papers* 5: 249–263.
1955. Concerning marginal utility. *Oxford Economic Papers* 7: 170–176.
1958. Utility and the 'ordinalist fallacy'. *Review of Economic Studies* 25: 172–181.

References

Armstrong, W.E., S. Daniel, and A.A. Francis. 1974. A structural analysis of the Barbados economy, 1968, with an application to the tourist industry. *Social and Economic Studies* 23: 493–520.
Georgescu-Roegen, N. 1954. Choice, expectations and measurability. *Quarterly Journal of Economics* 68: 503–534.
Liep, J. 1983. Ranked exchange in Yela (Rossel Island). In *The Kula*, ed. J.W. Leach and E.R. Leach. Cambridge: Cambridge University Press.
Little, I.M.D. 1950. The theory of consumer's behaviour: A comment. *Oxford Economic Papers* 2: 132–136.
Malinowski, B. 1921. The primitive economy of the Trobriand Islanders. *Economic Journal* 31: 1–16.
Malinowski, B. 1922. *Argonauts of the Western Pacific*. London: Routledge & Kegan Paul.
Mauss, M. 1925. *The gift*. London: Routledge & Kegan Paul.
Ng, Y.-K. 1975. Bentham or Bergson? Finite sensibility, utility functions and social welfare functions. *Review of Economic Studies* 42: 545–569.
Urry, J. 1985. W.E. Armstrong and social anthropology at Cambridge. *Man* 20: 412–423.

Arndt, Heinz Wolfgang (Born 1915)

R. D. Freeman

Keywords
Arndt, H. W; Indonesia; Keynesianism; Planning

JEL Classifications
B31

Born February 1915 in Breslau, Germany (now Wroclaw, Poland), Arndt was educated at Oxford University (1933–8) and London School of Economics (1938–41). After two years as a research assistant at the Royal Institute of International Affairs, Arndt was Assistant Lecturer in Economics, University of Manchester (1943–6), Senior Lecturer, University of Sydney (1946–50) and then Professor of Economics in the School of General Studies and Research School of Pacific Studies, Australian National University (1951–80). He became Emeritus Professor of Economics, Australian National University, in 1981. His many prestigious appointments include Member, Governing Council, United Nations Asian Institute for Economic Development and Planning (1969–75); Deputy Director, OECD (1972) and Chairman, Expert Group on Structural Change and Economic Growth Commonwealth Secretariat (1980).

Arndt first came to prominence in 1944 with his analytical economic analysis of the interwar period in which he argued the structuralist thesis that market forces could not correct the existing major disequilibria in the world economy. He recommended cooperative planning in the post-war period, involving controls on the volume and directions of international trade and investment and international cooperation if not supranational economic authorities.

His major contributions were in policy-oriented economic research with particular reference to developing countries in the Pacific Basin. A leading authority on the Indonesian economy as well as other Asian economies, led Arndt to start the *Bulletin of Indonesian Economic Studies* in 1965; he was also instrumental in the establishment in the Australian National University of a major research school on Asian economic development.

A prolific writer, Arndt was an important influence in Australian academic and policy circles in developing post-war understanding and acceptance of Keynesian macroeconomic analysis.

Selected Works

1944. *The economic lessons of the nineteen-thirties.* Oxford: Oxford University Press. Reprinted London: Frank Cass & Co., 1963; Italian trans., 1949; Japanese trans. 1978.

1954. A suggestion for simplifying the theory of international capital movements. *Economia Internazionale* 7, 469–481.
1955. External economies in economic growth. *Economic Record* 31(61), 192–214.
1957. *The Australian tading banks.* Melbourne: Cheshire 2nd edn, 1960; 3rd edn (with C.P. Harris), 1965; 4th edn (with D.W. Stammer), 1973; 5th edn (with W.J. Blackert), Melbourne: Melbourne University Press, 1977.
1978. *The rise and fall of economic growth*: *A study of contemporary thought.* Melbourne: Longman Cheshire.
1979. The modus operandi of protection. *Economic Record* 55(149), 149–155.
1981. Economic development: A semantic history. *Economic Development and Cultural Change* 29, 457–466.
1985. *A course through life.* Canberra: Australian National University.

Arrears

Mark E. Schaffer

Abstract

Arrears, in both common and general economic parlance, are overdue payments of any sort. The comparative economics literature has focused on the large-scale arrears of all sorts that emerged when central and eastern Europe and the former USSR began the transition to market economies. Soft budget constraints have been invoked in explanations of the growth of overdue trade credit or 'inter-enterprise arrears' in early transition, and in analyses of arrears to banks and tax arrears; studies of wage arrears in transition economies have focused on differential impacts across workers and firms and on weak institutions.

Keywords

Arrears; Inter-enterprise arrears; Soft budget constraint; Tax arrears; Trade credit arrears; Transition economies; Wage arrears

JEL Classifications

P3

Arrears, in both common and typical economic parlance, are overdue payments of any sort. In its last previous appearance in this dictionary, *Palgrave's Dictionary of Political Economy,* the term is defined simply as 'sums remaining unpaid after they are due' (Higgs 1925, p. 58). The context is usually one in which a payment is required by a contract or by law; hence the cross-references in the 1925 *Palgrave* entry to 'law of contract' and 'wages'. The same is true of contemporary usage: Internet search engines at the time of writing indicate the most commonly used term by far is 'mortgage arrears', followed by 'wage arrears' and 'tax arrears'. In most of economic science, arrears are generated by some behaviour or event, and it is the latter which is typically the focus of analysis. The analytical framework varies hugely with the object of analysis, and there is no theme that unites, for example, the analysis of consumer debt arrears and that of sovereign debt arrears.

The main exception to this, and the reason 'arrears' has reappeared in the *New Palgrave,* is the arrears phenomenon that emerged on a large scale when the countries of central and eastern Europe and the former USSR abandoned the socialist economic system and began the transition to market economies. The arrears phenomenon in transition economies arose when firms accumulated non-payments of obligations to various creditors, often on a very large scale. The natural way to analyse this phenomenon is to distinguish between the main categories of creditors to the firms that have accumulated arrears – other firms, banks, the state and employees – and between stocks and flows, late payments and non-payments.

In the comparative and transition economics literature, overdue debts of firms to other firms

has often been termed 'inter-enterprise arrears', though a more standard term from mainstream economics would be 'trade credit arrears'. The rapid emergence of large volumes of overdue trade credit in many formerly socialist countries in the early phase of the transition (1989–93) took many economists in both the policy and academic communities by surprise. In retrospect, this surprise partly reflects the fact that trade credit is an understudied phenomenon in general. After early, rapid growth, the volumes of both total trade credit and overdue trade credit in the transition economies stabilized at levels similar to those found in developed market economies – the equivalent of roughly 20–40 per cent and 10–20 per cent of GDP, respectively (Schaffer 1998). The eventual stabilization at levels found in normal market economies implies an approximate matching of inflows and outflows, and follows partly from the fact that late payment of trade credit is an endemic problem in market economies generally, as a reading of the business press and reports by factoring agencies will confirm. It also implies that firms in transition economies, including state-owned firms that had previously been unexposed to market forces, learned fairly rapidly to impose hard budget constraints on each other.

The early phase of rapid growth of trade credit arrears is a somewhat different matter. First, the payment systems that were used in socialist economies were typically very inflexible. Ickes and Ryterman (1992) argue that in Russia, the most studied country case of trade credit arrears, the combination of a lack of liquidity following price liberalization in January 1992 and a first-in-first-out (FIFO) queuing system for clearing payments generated 'payments gridlock' and thus rapid growth in arrears on payments to suppliers. The government's response in mid-1992 was to abandon the payment queuing system and, separately, to try to clear the accumulated backlog of payments with an accompanying injection of credit, amounting to a bailout of the enterprise sector. Second, the model of Perotti (1998) suggests that collusive non-payment by the enterprise sector can force a government bail-out via a 'too-big-to-fail' mechanism. Both explanations are examples of soft-budget constraints in action. This early phase of rapid growth also took place in the moderate- to high-inflation environments that followed price liberalization in these countries. The effective interest rate subsidy that accompanied trade credit thus involved a substantial discount to buyers, though it has also been suggested that sellers anticipated both inflation and payment delays, and incorporated a corresponding markup in their prices.

Arrears of firms to banks in transition economies is the phenomenon that is least specific to the transition experience. The large bad-debt problems that emerged following the start of transition have been analysed in the literature using the standard frameworks and tools for analysing systemic banking-sector problems. The limited evidence from these economies suggests that connected lending and directed state credits became a primary mechanism in the slower reformers for bailing out firms and softening budget constraints into the 1990s and beyond. Large-scale tax arrears of firms, by contrast, are peculiar to the transition experience. In developed market economies, tax arrears of firms are a phenomenon largely associated with exit of insolvent firms, and the scale is relatively small; New Zealand has been cited as an example, with a stock of tax arrears amounting to one or two percentage points of GDP, and annual write-offs of uncollectible taxes coming to less than one-half of one percentage point of GDP. In the first five or ten years of transition, however, available evidence suggests that government toleration of non-payment of taxes was common even in the more rapidly reforming countries. Rough estimates of the scale of tax arrears range from two to 12 percentage points of GDP for the stock, and one to seven percentage points for the annual flow (Schaffer 1998), and the empirical evidence suggests they were one of the main mechanisms governments used to soften the budget constraints of firms.

Lastly, large-scale and persistent wage arrears are also peculiar to transition economies, though in this case mostly limited to the countries of the former USSR. The scale of the wage arrears of firms at their peak – in aggregate, several percentage points of GDP – was typically smaller than trade credit and even tax arrears, but substantial in

comparison with monthly wages. Payment of wages to employees several months in arrears was commonplace, and the absence of indexation imposed an extra cost in the high-inflation period of the early 1990s and following the burst of inflation that accompanied the collapse of the rouble in mid- 1998. Wage arrears have sometimes been an important adjustment mechanism for labour markets in transition economies, partially absorbing negative shocks that would otherwise be fully reflected in actual wages or employment levels. The empirical evidence suggests that most wage arrears were late payments rather than non-payments, and with important distributional impacts with respect to household income. The social consequences of uncertainty and irregularity of wage payments were substantial, since workers in these countries had limited savings to fall back on and even less access to consumer credit markets, and thus faced great difficulties in smoothing income. Patterns across firms and workers in wage arrears have been related to firm, worker, and economy-wide characteristics (state-owned, poorly performing firms; workers in rural areas, outside options; tight credit policies; workers in sectors such as health and education, funded by the government budget), and to weak institutional environments that made it possible for firms to violate wage contracts at relatively low cost (see, for example, Lehmann et al. 1999; Earle and Sabirianova 2002).

See Also

- ▶ Assets and Liabilities
- ▶ Soft Budget Constraint
- ▶ Transition and Institutions

Bibliography

Earle, J.S., and K.Z. Sabirianova. 2002. How late to pay? Understanding wage arrears in Russia. *Journal of Labor Economics* 20: 661–707.

Higgs, H. (ed.). 1925. *Palgrave's dictionary of political economy*. London: Macmillan and Co.

Ickes, B.W., and R. Ryterman. 1992. The interenterprise arrears crisis in Russia. *PostSoviet Affairs (Formerly Soviet Economy)* 8: 331–361.

Lehmann, H., J. Wadsworth, and A. Acquisti. 1999. Grime and punishment: Job insecurity and wage arrears in the Russian Federation. *Journal of Comparative Economics* 27: 595–617.

Perotti, E.C. 1998. Inertial credit and opportunistic arrears in transition. *European Economic Review* 42: 1703–1725.

Schaffer, M.E. 1998. Do firms in transition economies have soft budget constraints? A reconsideration of concepts and evidence. *Journal of Comparative Economics* 26: 80–103.

A

Arrow, Kenneth Joseph (Born 1921)

Ross M. Starr

Abstract

Kenneth Arrow is the author of key post-Second World War innovations in economics that have made economic theory a mathematical science. The Arrow Possibility Theorem created the field of social choice theory. Arrow extended and proved the relationship of Pareto efficiency with economic general equilibrium to include corner solutions and non-differentiable production and utility functions. With Gerard Debreu, he created the Arrow–Debreu mathematical model of economic general competitive equilibrium including sufficient conditions for the existence of market-clearing prices. Arrow securities and contingent commodities extend the model to cover uncertainty and provide a cornerstone of the modern theory of finance.

Keywords

American economic association; Arrow, K; Arrow–debreu model of general equilibrium; Arrow-pratt index of risk aversion; Arrows' theorem; Bergson social welfare function; Black, D; Condorcet, J.-A.-N; Constant-elasticity-of-substitution production function; Contingent commodities; Control theory; Convexity: in theorems of welfare economics; Corner solutions; Cowles commission for research in economics; Debreu, G; Discrimination: racial; Econometric society; Endogenous

growth; Expected utility theorem; First fundamental theorem on welfare economics; Fixed-point theorem; Game theory; Growth models; Hahn, F; Hicks, J; Hotelling, H; Impossibility theorem; Independence of irrelevant alternatives; Inventory policy: optimal; Knowledge: as externality; Kuhn-Tucker theorem; Marginal rate of substitution; Mathematical economics; Medical insurance; Moral hazard; Nash equilibrium; Pareto efficiency; Paradox of voting; Partial equilibrium theory; Principal-agent problem; In medical care; Racial discrimination; Arrow–pratt index of risk aversion; Russell, B; Samuelson, P; Second fundamental theorem on welfare economics; Securities markets; Social choice; Solow, R; Stigler, G; Tarski, A; Technical change; Tinbergen model; Tobin, J; Uzawa, H; Walras's law

JEL Classifications
B31

Kenneth Arrow is a legendary figure, with an enormous range of contributions to twentieth-century economics, responsible for the key post-Second World War innovations in economic theory that allowed economics to become a mathematical science. His impact is suggested by the number of major ideas that bear his name: Arrow's Theorem, the Arrow–Debreu model, the Arrow–Pratt index of risk aversion, and Arrow securities.

Four of his most distinctive achievements, all published in the brief period 1951–54, are as follows:

Arrow Possibility Theorem. Social Choice and Individual Values (1951a) created the field of social choice theory, a fundamental construct in theoretical welfare economics and theoretical political science.

Fundamental Theorems of Welfare Economics. 'An extension of the basic theorems of classical welfare economics' (1951b) presents the First and Second Fundamental Theorems of Welfare Economics and their proofs without requiring differentiability of utility, consumption, or technology, and including corner solutions (zeroes in quantities of inputs or outputs).

The Arrow–Debreu model of general economic equilibrium. 'Existence of equilibrium for a competitive economy' (with Gerard Debreu 1954) creates the mathematical model of a competitive economy. The article formalizes the cross-effects between markets (effect of one market's price on another's demand and supply) and provides sufficient conditions for the existence of prices allowing decentralized market-clearing general equilibrium of a market economy. This model is central to the study of markets and welfare economics; it is now a standard of the field.

Securities markets and risk-bearing. 'Le rôle des valeurs boursières pour la répartition la meilleure des risques' (1953) introduces the concept of a 'contingent commodity'. The article formalizes the role of markets, including financial markets, insurance and the stock market, in resource allocation; it is a cornerstone of the modern theory of finance.

Personal and Intellectual History

Kenneth Arrow was born in New York City on 23 August 1921. He describes his family circumstances as financially comfortable during the 1920s, but 'my father lost everything in the great depression and we were very poor for about 10 years . . . When it came to college, my family's poverty constrained me to attend the City College' (Breit and Spencer 1986, p. 45). Free tuition at City College of New York (CCNY) gave a generation of New Yorkers their start on success. The searing experience of the Depression affected career ambitions. Arrow thought he should pursue the safe career of a high-school mathematics teacher. He took education courses and he had a very successful period of practice teaching in mathematics, preparing students for the New York State Regents examination. However, the roster of applicants for New York City teachers' positions was already filled.

Arrow graduated from CCNY in 1940 with the unusual combination of a mathematics major and

a Bachelor of Science in Social Science. While at CCNY he studied with Alfred Tarski in a course on the calculus of relations. Arrow was a proof-reader for Tarski's *Introduction to Logic* (1941). He entered Columbia University for graduate study and received an MA in mathematics in June 1941. Harold Hotelling, a statistician with an appointment in the economics department, was the decisive influence. Arrow notes, 'When I took [Hotelling's] course in mathematical economics, I realized I had found my niche' (Breit and Spencer 1986, p. 45). With the inducement of a fellowship in economics, Arrow transferred to the economics department for the rest of his graduate study.

Arrow's graduate work at Columbia was interrupted by the Second World War. During the war Arrow was a weather officer in the US Army Air Corps achieving the rank of Captain, working in the Long Range Forecasting Group. Arrow's first published paper comes from that period, 'On the Use of Winds in Flight Planning' (1949a). The group's principal task was to forecast the number of rainy days in air combat areas – a month in advance. The young statisticians in the Weather Division subjected the prediction techniques in use to statistical test against a simple null hypothesis based on historical data. Finding that prevailing techniques were not significantly more reliable than the null, the junior officers sent a memo to the General of the Air Corps suggesting that the group be disbanded. Six months later, the General's secretary replied on his behalf: 'The general is well aware that your forecasts are no good. However, they are required for planning purposes.' The group remained intact.

In 1946 Arrow returned to graduate study at Columbia. Harold Hotelling had by then left for the University of North Carolina's newly formed statistics department. The concern about making a living persisted. Arrow considered a non-academic career as a life insurance actuary. Tjalling Koopmans (at a Cowles Commission meeting in Ithaca, New York) advised him that actuarial statistics would prove unrewarding, saying, with characteristic reticence, 'There is no music in it.' Fortunately for economic science, Arrow followed this advice and decided to continue a research career.

In 1947 Arrow joined the (now legendary – then fledgling) research group at the Cowles Commission for Research in Economics at the University of Chicago. It seemed a golden age – all the ideas of mathematical economic theory and econometrics were being newly discovered. The close friendships and collaborations among colleagues of the Cowles Commission lasted a lifetime. Arrow describes the setting as a 'brilliant intellectual atmosphere ... with eager young econometricians and mathematically inclined economists under the guidance of Tjalling Koopmans and Jacob Marschak' (Lindbeck 1992, p. 107).

Jacob Marschak, the Cowles Commission Research Director, arranged for the Commission to administer the Sarah Frances Hutchinson Cowles Fellowship for women pursuing quantitative work in the social sciences (the Fellowship had originally specified a preference that fellows be women of the Episcopal Church of Seneca Falls, New York [reported in conversation with Jacob Marschak]). The fellows were Sonia Adelson (subsequently married to Lawrence Klein) and Selma Schweitzer. Kenneth Arrow and Selma Schweitzer were married in 1947.

Graduate study 1946–50, through Columbia, Chicago, Cowles, RAND and Stanford, included a daunting search for a worthy dissertation topic. Prospects considered and rejected included revising and restating the Tinbergen model (Tinbergen 1939), and revising and restating Hicks's *Value and Capital* (1939). No topic seemed worthy. Then lightning struck: Arrow invented an entire field of economics with his dissertation 'Social Choice and Individual Values'. The Columbia Ph.D., with Professor Albert Hart as dissertation advisor, was granted in 1951. As an econometrician, T. W. Anderson of Columbia (subsequently Arrow's colleague at Stanford) was called upon to pass judgement on a draft thesis unrecognizable as economics to Ken's advisors; Anderson pronounced the work sound.

The summer of 1948 and several summers thereafter were spent at the recently formed RAND Corporation in Santa Monica, California,

a major centre of the newly emerging specialities of game theory and mathematical programming. In 1949 Arrow was appointed Acting Assistant Professor of Economics and Statistics at Stanford University, and rapidly became Professor of Economics, and of Statistics, with the eventual additional title of Professor of Operations Research. He moved to Harvard in 1968 (returning regularly to Stanford for summer workshops), and rejoined the Stanford faculty in 1979. He retired in 1991.

In the 1950s and 1960s at Stanford, economic theory and econometrics faculty and graduate students were located in Serra House (converted from the retirement residence of the first president of the university) under the auspices of the Institute for Mathematical Studies in the Social Sciences (IMSSS) organized under the leadership of Patrick Suppes. In his memorial remarks for his student, Walter P. Heller (1942–2001), Arrow describes the *esprit de corps*: 'Economic theory backed by serious mathematical reasoning was just beginning to be recognized. Our group of faculty and students in economic theory at Serra House. felt ourselves a community. Not an oppressed minority, but rather a vanguard. We were taking over!'

Stanford and UC Berkeley were centres of research in statistics and economic theory. The joint Berkeley–Stanford Mathematical Economics Seminar met biweekly at alternate campuses. The Berkeley group included Gerard Debreu, Roy Radner, Peter Diamond and Dan McFadden. Stanford's included Herbert Scarf and Hirofumi Uzawa. Uzawa came to Stanford on fellowship arranged by Arrow. Working on his own in Japan, he had written the manuscript eventually published as 'Gradient method for concave programming, II: Global stability in the strictly convex case' (Arrow et al. 1958a, ch. 7). It was a successful global stability analysis of gradient adjustment, following Arrow and Hurwicz's local analysis (available to Uzawa in manuscript, published in the same volume). Arrow read the manuscript and enthusiastically invited Uzawa to accept a fellowship at Stanford.

Although the profession is now used to mathematical expression, in the 1950s and 1960s the mathematical complexity of Arrow's work was regarded as forbidding. Although Arrow was the pre-eminent economic theorist at Stanford, he was not designated to teach in the required first-year graduate microeconomic theory course; it was presumed that the treatment would be excessively abstract for this general audience. His reputation for mathematical abstraction provided the excuse for a jest when Arrow received the 1957 John Bates Clark Award of the American Economic Association (presented to a leading economist under the age of 40). At the presentation ceremony, introductory remarks were made by George Stigler, who reportedly advised Arrow, in a stage whisper, 'You should probably say, "Symbols fail me".'

Under the administration of President J.F. Kennedy, Arrow and Robert Solow served on the research staff of the Council of Economic Advisers. That was a remarkable group: Walter W. Heller, chair, Kermit Gordon and James Tobin. The Council and its staff then included three future Nobel laureates: Arrow, Solow and Tobin.

Academic travels abroad included visits to the Institute for Advanced Studies in Vienna in the summers of 1964 and 1971, and productive years at Churchill College, Cambridge, in 1963–64 and 1970, for collaboration with Frank Hahn on *General Competitive Analysis* (1971a).

To no one's surprise, Arrow received the 1972 Nobel Prize in Economic Sciences (jointly with the distinguished British economic theorist, John Hicks of Oxford). Aged 51 at the time of the award, he is (at this writing) by far the youngest recipient of the Nobel Prize in Economics.

Testimony to Arrow's qualities as a dissertation advisor, a teacher of the next generation of economists, is abundant. The flurry of former students volunteering to contribute to the Festschrift by Heller et al. (1986) was overwhelming. The most personal tribute is the number of leading colleagues whose children have studied with Arrow. Jacob Marschak's son Thomas Marschak and Walter W. Heller's son Walter P. Heller wrote their doctoral dissertations with Arrow as principal advisor. Any list of Arrow's students (dissertation advisees, postdocs, and so forth) is a partial listing. They are numerous and are enthusiastically devoted to him, playing leading roles in academic and research economics. A selection

includes: Theodore Bergstrom (UC Santa Barbara), David Bradford (Princeton University), Michael Bruno (Hebrew University, Bank of Israel), Graciela Chichilnisky (Columbia University), Peter Coughlin (University of Maryland), John Geanakoplos (Yale University), Louis Gevers (Université de Namur, Belgium), John Harsanyi (UC Berkeley), Walter P. Heller (UC San Diego), Peter Huang (University of Minnesota Law School), Takatoshi Ito (University of Tokyo), Jean-Jacques Laffont (Université des Sciences Sociales, Toulouse, France), Robert Lind (Cornell University), Thomas Marschak (UC Berkeley), Eric Maskin (Institute for Advanced Study, Princeton), Roger Myerson (University of Chicago), Hajime Oniki (Osaka-Gakuin University, Osaka, Japan), Heraklis Polemarchakis (Brown University), Karl Shell (Cornell University), Ross Starr (UC San Diego), David Starrett (Stanford University), Nancy Stokey (University of Chicago), Laurence Weiss (Goldman Sachs Corp.), Ho-Mou Wu (National Taiwan University), and Menahem Yaari (Hebrew University, Jerusalem).

A range of stories depict Arrow as a legendary larger-than-life figure:

'Arrow is personally accessible and unpretentious, addressed as "Ken" by students, colleagues, and staff. . . Arrow thinks faster than he – or anyone else – can talk. Conversation takes place at such a rapid pace that no sentence is ever actually completed' (Heller et al. 1986, v. 1, p. xvii). The breadth of Arrow's knowledge is repeatedly a surprise, encompassing Chinese art, English history and the works of Shakespeare. At the 80th birthday celebration, Eric Maskin related the following example:

> On almost any subject arising in conversation, Arrow turns out to know a lot more than you do. Tired of being repeatedly shown up by their senior colleague, a group of junior faculty once concocted a plan. They first read up thoroughly on the most arcane topic they could think of – the breeding habits of gray whales. On the appointed day they gathered in the coffee room and waited for Ken to come in. Then they started talking about whales, concentrating on the elaborate theory of a marine biologist named Turner on how gray whales found their way back to the same breeding spot year after year. Ken was silent . . . they had him at last! With a sense of delicious triumph, they continued to discuss whales, and Ken looked more and more perplexed. Finally, he couldn't hold back: 'But I thought that Turner's theory was entirely discredited by Spencer, who showed that the hypothesized homing mechanism couldn't possibly work.'

Arrow's presence in seminars is distinctive. He may open his (copious) mail, juggle a pencil, seem inattentive. He will then make a comment demonstrating that he is several steps ahead of the speaker. He will make clear that the history of economic thought includes abundant antecedents (which he can readily cite from memory) for the issues under discussion.

Social Choice and Individual Values: The General Possibility Theorem

Social Choice and Individual Values was Arrow's doctoral dissertation, published as a Cowles Commission monograph. There are very few new ideas in economics. Arrow's General Possibility Theorem is as novel and fundamental as they come. The paradox of voting (cyclic majorities) appears to have been well-known, though not well formalized; Arrow (1951a) and Duncan Black (1948) both take it as understood. A review of the literature shows that it is attributable to Condorcet et al. (1785). The paradox – intransitivity of choice from majority vote based on voters with transitive preferences – can be stated simply.

Think of three voters trying to decide by majority vote among three possibilities, A, B and C. Each of the individual voters has transitive (rational) preferences. Voter 1 prefers A to B and prefers B to C. Voter 2 prefers B to C and C to A. Voter 3 prefers C to A and A to B. Then there is a majority of voters preferring A to B (voters 1 and 3), and a majority preferring B to C (voters 1 and 2). If group decision-making is also transitive (rational), then the group should prefer A to C. But just the opposite occurs; there is a majority preferring C to A (voters 2 and 3). Despite the transitivity of individual preferences, the group preference on pairs of alternatives, as expressed by majority vote, is intransitive (irrational).

Arrow's General Possibility Theorem (also known as 'Arrow's Theorem', the 'Arrow Possibility Theorem' or the 'Arrow Impossibility Theorem') shows that the paradox is not merely an anomaly but intrinsic to group decision-making. The theorem has been a focus of vigorous study for generations. An elegant proof in Sen (1986) is particularly striking since it is framed as a generalization of the Condorcet paradox.

The Possibility Theorem suggests four reasonable criteria for a group decisionmaking mechanism, all of which are fulfilled by majority voting (assume at least three possible choices and at least three voters):

1. *Unrestricted Domain*. The decision-making mechanism can accommodate all logically possible preferences on the available choices.
2. *Pareto Principle*. If everyone prefers one alternative over another, the group decision should have that preference as well.
3. *Independence of Irrelevant Alternatives*. In choosing between any two alternatives, group decision-making takes account only of individual preferences on those alternatives; preferences on a third possibility do not enter the choice between those two.
4. *Non-dictatorship*. There is no single person whose preferences will always be followed by the group decision-making mechanism.

The Possibility Theorem says that no decision-making mechanism that fulfils all four of the above conditions results in transitive (rational) group choices based on transitive (rational) individual preferences. The Condorcet paradox is not merely an anomaly. It is unavoidable. It represents a fundamental defect in group decision-making.

Each of the four above conditions is essential to the theorem; there are examples of transitive group decision-making mechanisms that fulfil any three but not four. Of the four, the most controversial is Independence of Irrelevant Alternatives; it prevents voluntary misstatement by a voter of his preferences from being an attractive strategy (overstating dislike of a third option to make a preferred one of two succeed in a weighted voting scheme).

At the time *Social Choice and Individual Values* was published, the logic of group decision-making was not even recognized as an economic issue. Since then there has been an overwhelming blossoming of the 'social choice' field. It is a topic for the *Handbook of Mathematical Economics* (Sen 1986); thousands of journal articles deal with it; every graduate student in economics is introduced to it. Kenneth Arrow created the field by formalizing a result that says the object of the field is unachievable.

The book also had a significant impact in a second direction: treating economic theory as an axiomatic logical field rather than as a sphere of calculation. *Social Choice* was one of the first essays, certainly the first monograph, to treat economics with the same generality and logical rigour as classical geometry. This approach was to be repeated in the next of Arrow's several major works in general equilibrium theory and classical welfare economics.

How did Arrow come to develop this structure? It was during the first summer, in 1948, at RAND that several strands of thought came together. The Condorcet paradox of cyclic majorities was common knowledge (though not the attribution to Condorcet). Independently of Duncan Black (1948), Arrow developed the restriction of individual preferences to the single-peaked format as a solution, but then realized that he'd been scooped when he read Black's result in the *Journal of Political Economy*. He was aware of the ambiguity in describing the optimizing policy of a business firm under uncertainty: profit maximization is no longer well-defined and majority voting of shares is subject to the Condorcet paradox. Arrow's techniques of logical formalization were ready. As a high-school student he had read Russell's *Introduction to Mathematical Philosophy* (1920); at CCNY he became familiar with Tarski's *Introduction to Logic* (1941) and the calculus of relations. With that preparation, it was obvious that the indifference curve approach used by economists was a form of a logical ordering. Axiomatic treatment came naturally.

RAND was the centre of the developing field of game theory, which was being used to formalize discussions of strategic behaviour in

international relations. During a coffee break the logician Olaf Helmer posed the following problem. Game theory supposes rational strategic behaviour among optimizing agents. The maximand of an individual may be well-defined, perhaps as a utility function; but what is the maximand of a country? Arrow replied that a Bergson social welfare function should represent a country's maximand. That set him to work. Demonstrating that his answer to Helmer was fundamentally and necessarily inadequate is the meaning of the Possibility Theorem. Arrow started the inquiry by looking at a variety of group decision-making mechanisms. They all looked wrong; either they led to intransitivity or they violated the Independence of Irrelevant Alternatives, so that preferences for an alternative that was out of the running nevertheless entered the group's decision. He was led to formalize the conditions of group decision-making, reflecting a long-standing interest in axiomatic reasoning. 'The development of the theorems and their proofs then required only about 3 weeks, although writing them as a monograph took many months' (1983a, p. 4).

Extension of the Fundamental Theorems of Welfare Economics

In the 1940s welfare economics in mathematical form (the relationship of market equilibrium to economically efficient allocation) was very much a matter of the calculus (Samuelson 1947). Marginal rates of substitution (ratios of marginal utilities) were equated to marginal rates of transformation (ratios of marginal products of factors) which were equated to price ratios. This is a sound viewpoint so long as the underlying functions are differentiable and the quantities of goods and factors are in a range where they can be varied. Arrow's view was that there is a fundamental weakness to this approach in the presence of non-negativity constraints on quantities. It works only when quantities are strictly positive. That is, the calculus doesn't treat corner solutions. But almost every practical economic solution is a corner solution: it is rare to find that all quantities of all possible goods and all possible inputs are used in strictly positive quantities. This is particularly true when differing qualities or varieties of similar goods are treated distinctly (white, sourdough and rye breads are distinct commodities, as are luxury and efficiency apartments). There must be a welfare economics that includes corner solutions; it must be possible to present welfare economics without the calculus.

Arrow attributes his insight to a seminar presentation on the fundamental theorems of welfare economics given by Paul Samuelson at the University of Chicago, in Samuelson's style using the calculus (1983b, p. 14). The diagrams that illustrated the equations depicted a separating hyperplane. Arrow had learned of the fundamental role of convexity and the separating hyperplane theorem at RAND in the summer of 1948. The result of these reflections is 'An extension of the basic theorems of classical welfare economics' appearing in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The conference was held in the summer of 1950 in Berkeley, and the proceedings appeared a year later. There, the First and Second Fundamental Theorems of Welfare Economics are stated in terms of real analysis and convex sets, without the use of the calculus and including corner solutions.

At the level of the firm and the household, characterizing optimizing behaviour at corner solutions is the job of the Kuhn–Tucker Theorem. In a case of simultaneous discovery of related ideas, that theorem was first publicly presented at the same Berkeley Symposium (Kuhn and Tucker 1951).

First Fundamental Theorem of Welfare Economics: Every competitive equilibrium allocation is Pareto efficient. This result does not require convexity of tastes or technology, though convexity may be useful in establishing the existence of equilibrium prices.

Second Fundamental Theorem of Welfare Economics: In an economy with convex technology and preferences, every Pareto-efficient allocation can be sustained as a competitive equilibrium with appropriate prices subject to a redistribution of ownership shares in firms

and redistribution of endowment (except that some low-income households may be expenditure minimizers subject to utility constraint, rather than utility maximizers subject to budget).

Neither of these results depends on positivity of quantities or on differentiability of the functions or relations. The generality of the results, the use of a formal mathematical structure of assumptions, theorems and proofs was again novel. It meant that economics was becoming closer to formal mathematics.

General Equilibrium Theory

In the early 1950s, Arrow (at Stanford) pursued, largely by correspondence, joint work on general equilibrium theory with Gerard Debreu, who was then at the Cowles Commission in Chicago. The theory of general economic equilibrium recognizes that the economy is an interactive system. Decisions and prices in one market have a direct impact on supply and demand in other markets. The question Arrow and Debreu treated is: under what (sufficiently general and formalized) conditions can there be prices so that all markets simultaneously clear? This issue is known as 'the existence of economic general equilibrium'. The term 'general' equilibrium refers to the many markets simultaneously clearing, as opposed to 'partial' equilibrium where a single market is considered in isolation. Moreover, the theory allows – or forces – the theorist to formulate relatively complete models of the economy. The result of these inquiries has been an intellectual revolution and an intellectual foundation for market economics. A half-century after it was introduced to economics, the Arrow–Debreu model is the cornerstone and workhorse of our theory of markets and resource allocation.

Abraham Wald, with whom Arrow had studied at Columbia, had written several papers in the field (while in Vienna in the 1930s before emigrating to avoid the Nazi takeover) but had run up against fundamental mathematical difficulties (Wald 1934–35, 1936). He explained to Arrow that the problem was 'very difficult', advice that was enough to discourage the young economic theorist for some years. It was the recognition by Arrow and Debreu of the importance of using a fixed point theorem that led to major progress in this area. (Credit for independent discovery of the importance of fixed point theorems in this context is due to Lionel McKenzie 1954. The use of a fixed point theorem for demonstrating the existence of an equilibrium [of a game] was pioneered by John Nash 1950. See Debreu 1983).

Arrow describes his early thoughts on the subject and the interaction with ideas current at the time (particularly the Nash equilibrium of N-person games) thus:

> My original approach, for what it is worth, was to formulate competitive equilibrium as the equilibrium of a suitably chosen game. The players of this fictitious game were the consumers, a set of 'anticonsumers' (one for each consumer), producers, and a price chooser. Each consumer chose a consumption vector, each anticonsumer a nonnegative number (interpretable as the marginal utility of income), each firm a production vector, and the price chooser a price vector on the unit simplex. The payoff to a consumer was the utility of his consumption vector plus the budgetary surplus (possibly negative, of course) multiplied by the anticonsumer's chosen number. The payoff to an anticonsumer was the negative of the payoff to the corresponding consumer. The payoff to the firm was profit and to the price chooser the value of excess demand at the chosen prices. This is a well-defined game. The existence of equilibrium does not follow mechanically from Nash's theorem, since some of the strategy domains are unbounded.
>
> Debreu and I sent our manuscripts to each other and so discovered our common purpose. We also detected the same flaw in each other's work; we had ignored the possibility of discontinuity when prices vary in such a way that some consumers' incomes approach zero. [The possibility of discontinuity in demand at incomes where household consumption is on the boundary of the possible consumption set is known as the 'Arrow corner'.]. We then collaborated, mostly by correspondence, until we had come to some resolution of this problem. In the main body of the work we followed more closely Debreu's more elegant formulabased on the concept of generalized games, which eliminated the need for 'anticonsumers.' (1983b, pp. 58–9)

The papers of Arrow and Debreu (1954) and McKenzie (1954) were presented to the 1952 meeting of the Econometric Society. Publication

of 'Existence of equilibrium for a competitive economy' represents a fundamental step in the revision of economic analysis and modelling, demonstrating the power of a formal axiomatic approach with relatively advanced mathematical techniques. The approach of the field is revolutionary: it fundamentally changes our way of thinking. Once we see things this way, it is hard to conceive of them otherwise.

Sufficient conditions for the existence of market-clearing prices – consistent with one-another – for N distinct commodities are: (a) demand and supply are continuous as a function of prices, and (b) Walras's Law. These properties are derived from fundamental assumptions on the structure of preferences and endowments of households and the technology of firms. The theory is general enough to include point-valued and (convex) set-valued demand and supply.

Debreu's *Theory of Value* (1959) made the Arrow–Debreu general equilibrium model accessible to the wider profession. The implications for economic theory as a discipline were multifaceted: *general* equilibrium, treating all markets as interacting together, became systematic; the axiomatic method was set firmly in place as part of economic theory. Economic theory could be as precise and logically demanding as geometry. The potential of formal theory to generalize could be brought to bear. The Arrow–Debreu treatment proved, with full mathematical rigour, that any economy fulfilling the model's clearly and generally specified assumptions would produce its specified results.

A number of articles (principally co-authored with Leonid Hurwicz, 1958b, 1959) treat the stability of general equilibrium. Though Arrow and Debreu (1954) establishes the existence of market clearing prices, it does not derive 'equilibrium' as the rest point of a dynamic system. The stability question focuses on how a price adjustment system will lead to market clearing prices. Since prices in each market (at least potentially) enter into the excess demands of all markets, there is plenty of room for price adjustments to go awry. This body of literature sorts out and proves sufficient conditions for adjustment to be successful. Bottom line: a sufficient condition is that other markets do not excessively interfere with excess demands on any single market; if the principal determinant of excess demands for each good is the price of that good, then price adjustment to market clearing will be successful.

The effect of the introduction of the Arrow–Debreu model on economic theory has been overwhelming. Every graduate-level textbook in microeconomic theory discusses it. Whole classes of economic theorists describe their speciality as 'general equilibrium theory'. In the 15 years following publication of *Theory of Value*, a major focus of pure theory was understanding and extending the model. This included its relationship to bargaining (Debreu and Scarf 1963), to large economies (Aumann 1966) and to computing general equilibrium prices (Scarf and Hansen 1973). It was further elaborated by Arrow and Hahn (1971a).

Contingent Commodities

Part of the power of mathematics is generalization. If you've solved a problem once, you don't have to solve it again – even in different circumstances if you can show that the previous treatment applies. This was the brilliantly simple insight in the creation of the concept of 'contingent commodity'.

Arrow's thought had been influenced by Hicks's *Value and Capital*, including understanding the power of defining a commodity to include specification of time and location, and by L.J. Savage's lectures on mathematical statistics at Chicago, including a notion of the 'state of the world' as defining a random variable. (The 'state of the world' concept for defining a random variable is attributable to Kolmogorov [1933]). It was a fundamental step to combine these notions so that a commodity might be defined by what it is, where and when deliverable, *and by the 'state of the world' in which it is deliverable*.

By redefining a 'commodity' in this way as a 'contingent commodity', the complete structure of the Arrow–Debreu model of general equilibrium and economic efficiency could be applied. This is now typically described in the literature as

'a full set of Arrow–Debreu futures contracts'. The concept of an efficient (or 'optimal') allocation of risk-bearing is immediately evident as a consequence of the modelling structure. The next step is to suggest a security contract contingent on the state of the world payable in money – to economize on the number of actively traded commodities – now known as an 'Arrow security' or 'Arrow insurance contract'. This has been an extremely powerful concept, allowing researchers to formulate their ideas clearly; the Arrow security is a staple of twenty-first century theoretical finance.

The paper 'Le rôle des valeurs boursières pour la répartition la meilleure des risques', originally written in English, was translated into French for a conference at Centre National de Recherche Scientifique, Paris, in June 1952. Other conference participants included Jacob Marschak, Maurice Allais, L.J. Savage, Milton Friedman and Pierre Massé. It was published in French in *Econométrie* and the original English version appeared (as a 'translation') a decade later in *Review of Economic Studies*, after the notions had been introduced to English-speaking readers in *Theory of Value*.

Individual Behaviour Towards Risk, Economics of Medical Care, Learning by Doing

Treatment of uninsurable risk (where contingent commodities and Arrow securities are not available or correctly priced) has been a focus of Arrow's work for decades. It appears in the *Collected Papers*, the *Aspects of the Theory of Risk Bearing (Yrjo Jahnsson lectures)* (1965a), and in *Essays in the Theory of Risk Bearing* (1971b). These essays provide for many readers the most systematic treatment available of the statement and proof of the Expected Utility Theorem, derivation of the Arrow–Pratt risk aversion index, and a systematic framework for considering decision-making in an uncertain world.

Several papers (1963, 1965b) treat the economics of medical care, a setting where uncertainty, information as a scarce resource, and insurance all play a part. An element of the contribution is to state the issues in an abstract analytic economic framework. This reminds economists of why these problems are not textbook economics, and reminds non-economists that the economics textbook is useful. The historical setting in which these articles were written is pre-1990, that is, before health maintenance organizations (HMOs) became popular, when the principal form of medical insurance available was fee for service. They contain several insights (probably not unique to or first from Arrow, but effectively presented). For example, medical needs are uncertain so medical insurance is not merely a form of payment but is a response to risk. Again, medical insurance reduces the marginal cost of care as seen by the patient below actual cost, encouraging increased use (moral hazard consequence of insurance). Finally, medical care is distinct (but not unique) among commodities in that the decisions to incur care and the form that it should take are made to a large extent by the provider (the medical doctor) who is paid for providing care rather than by the buyer (patient). There is a resulting conflict of interest and reliance on professional norms. Arrow's treatment of the doctor-patient relationship as a seller-buyer interaction is an early appearance in the literature of the conflict we now recognize as the 'principal-agent problem' with an attendant family of issues.

In the eighteenth century Adam Smith noted that one of the benefits of specialization in production was that workers at specialized tasks learned how most effectively to perform them. Arrow's 'The economic implications of learning by doing' (1962) reflects in part the temper of the time – economic growth and growth models were a principal focus of theory and policy. In addition, it is a leap several decades ahead in growth theory. In contrast to growth models in the 1960s, it presents endogenous growth, a research topic that became an active focus decades later (Romer 1994). The study brings together two apparently disparate strands of economic modelling: technical change and the theory of external effects. The benefits of production in a particular line of work include not only output but the greater experience

of the firm and the workforce in production. Through production, workers and firms learn how to produce more with fewer inputs. To the extent that this knowledge is inappropriable or non-marketable, it provides an external benefit to the economy. This on-the-job experience will typically be under-provided relative to an economically efficient allocation.

Optimal Programming, Control Theory, Mathematical Statistics, Racial Discrimination, and the CES Production Function

In 16 books (not including the *Collected Papers*) and 250 technical articles, there are significant contributions to a breadth of issues in economics, mathematical programming and public policy. There's even some mathematical statistics (with Blackwell and Girshick 1949b).

One of the most useful – to other economists – is 'Capital-labor substitution and economic efficiency' by Arrow et al. (1961). It introduced the constant-elasticity-of-substitution (CES) production function, spawning an immense empirical literature.

Public Investment, the Rate of Return, and Optimal Fiscal Policy and several papers with Mordecai Kurz (1970) introduced control theory to the theory of the firm, to the theory of the household, and to public finance. A variety of books and articles treat mathematical programming and optimal inventory policy.

Several papers formally model racial discrimination in employment (1973). This is a tricky problem, and not merely because it is politically controversial. Pure microeconomic theory would suggest that there should be no racial discrimination by rational profit-maximizing employers; significant discrimination should result in below-market wage rates for the discriminated-against workers with resultant extra incentive for employers to hire them. How then can an economic model of optimizing behaviour explain the prevalence of racial discrimination? The answers (based on the racial views of employers, employees, customers) provide clues to locating the points of leverage that may lead to amelioration or policy.

What Have We Learned?

Arrow, along with Debreu, was a decisive figure in introducing the axiomatic method to economic theory. *Social Choice and Individual Values* and 'Existence of equilibrium for a competitive economy' fundamentally changed the agenda of economic theory. Formal logical reasoning and formal statement of assumptions and conclusions became the standard of pure theory (Suppes 2005). The axiomatic method need not be a straitjacket. Arrow's less formal work demonstrates the role of insight: observing actual economic activity and asking 'why?', where the acceptable class of answers reflects underlying principles of economic analysis. The result is a rich understanding of the nuance and power of economics.

Celebrations

Dedicated colleagues and students have done their best to show adulation and gratitude to Arrow. There has been a succession of public celebrations.

On Arrow's 65th birthday in August 1986, an immense birthday conference and party, known as the 'Arrowfest', took place at Stanford. It reunited colleagues and students from all over the world. There were two days of conference papers and testimonial remarks. A three-volume Festschrift was presented (on time) (Heller et al. 1986), including papers by 35 of Arrow's students and colleagues. Among the contributing authors were three (eventual) Nobel laureates: John Harsanyi, Amartya Sen and Robert Solow. The observance included a gala dinner with testimonial remarks and an expression of thanks from Arrow.

To observe his 70th birthday, the celebration was at the doctoral alma mater, a conference and social gathering in October 1991 titled 'Columbia Celebrates Arrow's Contributions'. The Festschrift volume (Chichilnisky 1999) included papers by 22 colleagues and students. The 70th

birthday was also the occasion of formal retirement from active faculty status at Stanford. That rite of passage was observed with a reception, including testimonials from colleagues, among them the senior colleagues who had been clever enough to recruit Arrow to Stanford two generations earlier. Stanford's Arrow Lecture Series was initiated, annually inviting distinguished speakers in economic theory in Arrow's honour.

A 40th anniversary party for general equilibrium theory was held in June 1993 at Center for Operations Research and Econometrics (CORE) of the Université Catholique de Louvain in Louvain-la-Neuve, Belgium. For several days and nights hundreds of professors, researchers and students from around the world presented papers, discussions and reminiscences of the speciality they had pursued for years. At the centre of the celebration were the twentieth-century founders of the field, Kenneth Arrow, Gerard Debreu and Lionel McKenzie.

There was a happy coincidence in 2001, when the 50th anniversary of *Social Choice and Individual Values* approximately coincided with Arrow's 80th birthday. A panel discussed the book's impact over the previous half century: Pat Suppes (Stanford University) on philosophy, John Ferejohn (Stanford University) on political science, and Eric Maskin (Institute for Advanced Study) on economics. The gathering included Professor Ted Anderson, who was at Columbia when *Social Choice* was submitted as Arrow's dissertation.

A dinner that evening featured moving toasts of appreciation by colleagues from around the world and presentations by Arrow's sons, Andy and David. The conclusion – sending the audience out singing into the evening – was the ad hoc musical group, the Economy Singers, singing advice to rising young economists: 'Brush Up Your Arrow, Start Quoting Him Now.'

To many students and colleagues, Kenneth Arrow is a source of inspiration and a focus of friendship and respect:

> . . . an inspirational teacher and colleague . . . The intellectual standards he set and the enthusiasm with which he approaches our subject are surely part of all of us. Those of us who have had a chance to know him well are particularly fortunate. We are far richer for the experience. (Heller et al. 1986, vol. 1, pp. xi, xvii)

See Also

- ▶ Arrow's Theorem
- ▶ Debreu, Gerard (1921–2004)
- ▶ Equilibrium (Development of the Concept)
- ▶ General Equilibrium (New Developments)
- ▶ Sen, Amartya (Born 1933)
- ▶ Uncertainty

Selected Works

1949a. On the use of winds in flight planning. *Journal of Meteorology* 6: 150–9.

1949b. (With D. Blackwell and M. A. Girshick.) Bayes and minimax solutions of sequential decision problems. *Econometrica* 17: 213–44.

1951a. *Social choice and individual values*. New York: Wiley.

1951b. An extension of the basic theorems of classical welfare economics. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley/Los Angeles: University of California Press.

1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. *Économetrie. Colloques Internationaux du Centre National de la Recherche Scientifique* 11, 41–7. Translated into English as 'The role of securities in the optimal allocation of risk-bearing', *Review of Economic Studies* 31(1964): 91–6.

1954. (With G. Debreu) Existence of equilibrium for a competitive economy. *Econometrica* 22: 265–90.

1958a. (With L. Hurwicz and H. Uzawa.) *Studies in linear and non-linear programming*. Stanford: Stanford University Press.

1958b. (With L. Hurwicz.) On the stability of the competitive equilibrium. *Econometrica* 26: 522–52.

1959. (With H. Block and L. Hurwicz.) On the stability of the competitive equilibrium, II. *Econometrica* 27: 82–109.

1961. (With H. Chenery, B. Minhas and R. Solow.) Capital-labor substitution and economic efficiency. *Review of Economics and Statistics* 43: 225–50.
1962. The economic implications of learning by doing. *Review of Economic Studies* 29: 155–73.
1963. Uncertainty and the welfare economics of medical care. *American Economic Review* 53: 941–73.
1965a. *Aspects of the theory of risk-bearing*. Helsinki: Yrjö Jahnssonin säätiö.
1965b. Uncertainty and the welfare economics of medical care: reply (The implications of transaction costs and adjustment lags). *American Economic Review* 55: 154–8.
1970. (With M. Kurz.) *Public investment, the rate of return, and optimal fiscal policy*. Baltimore/London: The Johns Hopkins University Press.
1971a. (With F. Hahn.) *General competitive analysis*. San Francisco: Holden-Day; Edinburgh: Oliver & Boyd.
1971b. *Essays in the theory of risk-bearing*. Chicago: Markham; Amsterdam and London: North-Holland.
1973. The theory of discrimination. In *Discrimination in labor markets*, ed. O. Ashenfelter and A. Rees. Princeton: Princeton University Press.
1983a. *Collected papers of Kenneth J. Arrow, volume 1: Social choice and justice*. Cambridge, MA: The Belknap Press of Harvard University Press.
1983b. *Collected papers of Kenneth J. Arrow, volume 2: General equilibrium*. Cambridge, MA: The Belknap Press of Harvard University Press.
1984a. *Collected papers of Kenneth J. Arrow, volume 3: Individual choice under certainty and uncertainty*. Cambridge, MA: The Belknap Press of Harvard University Press.
1984b. *Collected papers of Kenneth J. Arrow, volume 4: The economics of information*. Cambridge, MA: The Belknap Press of Harvard University Press.
1985a. *Collected papers of Kenneth J. Arrow, volume 5: Production and capital*. Cambridge, MA: The Belknap Press of Harvard University Press.
1985b. *Collected papers of Kenneth J. Arrow, volume 6: Applied economics*. Cambridge, MA: The Belknap Press of Harvard University Press.
1991. The origins of the impossibility theorem. In *History of mathematical programming*, ed. J. Leenstra, A. Rinnooy Kan and A. Schrijver. Amsterdam/New York/Oxford/Tokyo: CWI Amsterdam, North-Holland.

Bibliography

Aumann, R. 1966. Existence of competitive equilibria in markets with a continuum of traders. *Econometrica* 34: 1–17.
Black, D. 1948. On the rationale of group decision-making. *Journal of Political Economy* 56: 23–34.
Breit, W. and Spencer, R., (eds.). 1986. Kenneth J. Arrow. *Lives of the laureates: Seven nobel economists*. Cambridge, MA/London: MIT Press.
Chichilnisky, G. (ed.). 1999. *Markets, information, and uncertainty: Essays in economic theory in Honor of Kenneth J. Arrow*. New York: Cambridge University Press.
Condorcet, J.-A.-N. de Caritat, marquis de. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: Imprimerie royale.
Debreu, G. 1959. *Theory of value*, Cowles Foundation Monograph No. 17. New York: Wiley.
Debreu, G. 1983. Mathematical economics at Cowles. Online. Available at http://cowles.econ.yale.edu/about-cf/50th/debreu.htm. Accessed 9 July 2005.
Debreu, G., and H. Scarf. 1963. A limit theorem on the core of an economy. *International Economic Review* 4: 235–246.
Heller, W., R. Starr, and D. Starrett (eds.). 1986. *Essays in Honor of Kenneth J. Arrow*, vol. 3. Cambridge: Cambridge University Press.
Hicks, J. 1939. *Value and capital*. Oxford: Clarendon.
Kolmogorov, A. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: J. Springer.
Kuhn, H., and A. Tucker. 1951. Nonlinear programming. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman. Berkeley: University of California Press.
Lindbeck, A. (ed.). 1992. *Nobel Lectures: Economic sciences, 1969–1980*. Singapore: World Scientific Publishing Co.
McKenzie, L. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22: 147–161.

Nash, J. 1950. Equilibrium points of N-person games. *Proceedings of the National Academy of Sciences of the USA* 36: 48–49.
Romer, P. 1994. The origins of endogenous growth. *Journal of Economic Perspectives* 8(1): 3–22.
Russell, B. 1920. *Introduction to Mathematical Philosophy*. London/New York: Allen and Unwin/Macmillan.
Samuelson, P. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Scarf, H., and T. Hansen. 1973. *The computation of economic equilibria*. New Haven: Yale University Press.
Sen, A. 1986. Social choice theory. In *Handbook of mathematical economics*, vol. 3, ed. K. Arrow and M. Intriligator. New York: Elsevier.
Suppes, P. 2005. The pre-history of Kenneth Arrow's social choice and individual values. Unpublished manuscript. Stanford: Center for the Study of Language and Information, Stanford University.
Tarski, A. 1941. *Introduction to logic*. New York: Oxford University Press.
Tinbergen, J. 1939. *Business cycles in the United States of America, 1919–1932*. Geneva: Economic Intelligence Service, League of Nations.
Wald, A. 1934–35. Uberdie Produktionsgleichungen der Oekonomischen Wertlehre. *Ergebnisse eines Mathematischen Kolloquiums* 7, 1–6.
Wald, A. 1936. Uber einige Gleichungssystem der mathematischen Okonomie. *Zeitschrift fur Nationaloekonomie* 7, 637–70. Translated by Otto Eckstein as 'On some systems of equations of mathematical economics', *Econometrica* 19(1951) 368–403.

Arrow's Theorem

Kenneth J. Arrow

Abstract

Any satisfactory method of making a social choice should be in some measure representative of the individual criteria which enter into it, should use the range of possible actions, and should observe consistency conditions among the choices made for different data sets. Arrow's Theorem, or the Impossibility Theorem, states that there is no social choice mechanism which satisfies such reasonable conditions and which will be applicable to any arbitrary set of individual criteria. This article sets out the proof of the theorem.

Keywords

Arrow's th; Bentham J.; Compensation principle; Condorcet; Marquis de; Constitutions; Edgeworth F.; Hicks J.; Impossibility th; Independence of irrelevant alternatives; Intransitivity; Kaldor N.; Pareto principle; Preference orderings; Rawls J.; Scitovsky T.; Sidgwick H.; Single-peaked preferences; Social choice; Social welfare function; Voting

JEL Classifications

D7

Economic or any other social policy has consequences for the many and diverse individuals who make up the society or economy. It has been taken for granted in virtually all economic policy discussions since the time of Adam Smith, if not before, that alternative policies should be judged on the basis of their consequences for individuals; political discussions are less uniform in this respect, the welfare of an abstract entity, the state or nation, playing a role occasionally even in economic policy.

It follows that there are as many criteria for choosing social actions as there are individuals in the society. Furthermore, these individual criteria are almost bound to be different in some measure so that there will be pairs of policies such that some individuals prefer one and some the other. In the economic context, policies invariably imply distributions of goods, and in most policy choices, some individuals will receive more goods under one policy and others under the other. Individuals may also have different evaluations because of different concepts of justice or other social goals.

The individual criteria may be based on individual preferences over bundles of goods or individual preferences of a more social nature, with preferences over goods supplied to others. From the viewpoint of the formal theory of social choice, the criteria may even be judgements by others as to the welfare of individuals. The only assumption is that there is associated with each individual a criterion by which social actions are evaluated for that individual. Whatever their

origin, these criteria differ from individual to individual.

Every society has a range of actions, more or less wide, which are necessarily made collectively. Much of the debate on the foundations of social decision theory began with criteria for evaluating alternative tariff structures, including as the most famous illustration moving from a tariff to free trade. The redistribution of income through governmental taxes and subsidies provides another important case of an inherently collective decision which would be judged differently by different individuals.

If every individual prefers one policy to another, it is reasonable to postulate, as is always done by economists, that the first policy should be preferred. The problem arises of making social choices (between alternative collective policies) when some individual criteria prefer one policy and some another.

The fundamental question of social choice theory, then, is the following: given a range of possible social decisions, one of which has to be chosen, and given the criteria associated with the individuals in the society, find a method of making the choice. Not all methods of decision would be regarded as satisfactory. The method should be in some measure representative of the individual criteria which enter into it. For example, we would want the Pareto condition to be satisfied, that an alternative not be chosen if there is another preferred by all individuals. The method should use all the data, that is, both the range of possible actions and the individual criteria, and there are consistency conditions among the choices made for different data sets.

A pure case of social choice in action is voting, whether for the election to an office or a legislative decision. Here, the candidates or alternative legislative proposals are evaluated by each voter, and the evaluations lead to messages in the form of votes. The social decision, which candidate to elect or which bill to pass, is made by aggregating the votes according to the particular voting scheme used. The social decision then depends on both the range of alternatives (candidates or legislative proposals) available and the ranking each voter makes of the alternatives.

Voting procedures have one very important property which will play a key role in the conditions required of social choice mechanisms: only individual voters' preferences about the alternatives under consideration affect the choice, not preferences about unavailable alternatives.

Arrow's Theorem, or the Impossibility Theorem, states that there is no social choice mechanism which satisfies a number of reasonable conditions, stated or implied above, and which will be applicable to any arbitrary set of individual criteria.

Some terminology will be introduced in section 'The Language of Choice' of this entry. In section 'The Relevant Literature', there will be a brief review of the relevant literature as it was known to me prior to the discovery of the theorem. In section 'Statement of the Impossibility Theorem', I state the theorem with some variants and discuss the meaning of the conditions on the social choice mechanisms.

The Language of Choice

The formulation of choice and the criteria for it are those standard in economic theory since the 'marginalist revolution' of the 1870s as subsequently refined. There is a large set of conceivable *alternatives*; in any given decision situation, some given subset of these alternatives is actually available or feasible. This subset will be referred to as the *opportunity set.* Each individual can evaluate all alternatives. This is expressed by assuming that each individual has a *preference ordering* over the set of all alternatives. That is, for each pair of alternatives, the individual either prefers one to the other or else is indifferent between them (completeness), and these choices are consistent in the sense that if alternative x is preferred or indifferent to alternative y and y is preferred or indifferent to z, then x is preferred or indifferent to z (transitivity). This preference ordering is analogous to the preference ordering over commodity bundles in consumer demand theory. I have adopted the ordinalist viewpoint that only the ordering itself and not any particular numerical representation by a utility function is significant.

The *profile* of preference orderings is a description of the preference orderings of all individuals. For a given profile, the social choice mechanism will determine the choice of an alternative from any given opportunity set. In the case of an individual, it is assumed that the choice made from any given set of alternatives is that alternative which is highest on the individual's preference ordering. Analogously, it is assumed that social choices can be similarly rationalized. The social choice mechanism will have to be such that there exists a *social ordering* of alternatives such that the choice made from any opportunity set is the highest element according to the social ordering.

Therefore, a social choice mechanism or *constitution* is a function which assigns to each profile a social ordering.

The Relevant Literature

I will here review the literature on the justification of economic policy as I knew it in 1948–50. There was some work in economics and more in the theory of elections of which I was unaware, which I will briefly note.

The best-known criterion for what is now known as social choice was Jeremy Bentham's proposal for using the sum of individuals' utilities. Curiously, despite its natural affinity with marginal economics, it received very little serious use, possibly because its distributional implications were unacceptably extreme. Edgeworth applied the criterion to taxation (1925: originally published in 1897): see also Sidgwick (1901, ch. 7).

The use of the sum-of-utilities criterion required interpersonally comparable cardinal utility. A reluctance to make interpersonal comparisons led to the proposal of the compensation principle by Kaldor (1939) and Hicks (1939). Consider a choice between a current alternative x and a proposed chance to another alternative y. In general, some individuals will gain by the change and some will lose. The compensation principle asserts that the change should be made if the gainers *could* give up some of their goods in y to the losers so as to make the losers better off than under x without completely wiping out the gains to the winners. Notice that the compensation is potential, not actual. Since the only information used is the preference relation of each individual among three different alternatives, x, y, and a potential alternative derived from y by transfers of goods, no interpersonal comparisons are needed.

However, it turns out that the compensation principle does not define a social ordering. Indeed, Scitovsky (1941) showed that it was possible that the compensation principle would call for changing from x to y and then from y to x.

A different approach which sought to avoid not only interpersonal comparisons but also cardinal utility was the social welfare function concept of Bergson (1938). For each individual, first choose a utility function which represents his or her preference ordering. Then define social welfare as a prescribed function $W(U_1, \ldots, U_n)$ of the utilities of the n agents. For a given profile of preference orderings, if one of the utility functions is replaced by a monotone transformation (which represents therefore the same preference ordering), the function W has to be transformed correspondingly, so that social preferences defined by W are unchanged. In this formulation, a given social welfare function is associated with a given profile. There are no necessary relations among social welfare functions associated with different profiles.

It was also known to me, though I do not know how, that majority voting, which could be considered as a social decision procedure, might lead to an intransitivity. Consider three voters A, B and C and three alternatives, a, b and c. Suppose that A has preference ordering abc, B has ordering bca, and C the ordering cab. Then a majority prefer a to b, a majority prefer b to c, and a majority prefer c to a. Therefore, if we interpret a majority for one alternative to another as defining social preference, the relation is not an ordering. This paradox had in fact been discovered by Condorcet (1785), and there had been a small and sporadic literature in the intervening period (for an excellent survey, see Black 1958, Part II; also, Arrow 1973), but all

of this literature was unknown to me when developing the Impossibility Theorem.

There was one further very important paper, which I did know, the remarkable paper of Black (1948) on voting under single-peaked preferences. Suppose the set of alternatives can be represented in one dimension, for example, a choice among levels of expenditure (this was the case studied by Bowen (1943) who anticipated part of Black's results). Suppose individuals have different preference orders over the alternatives, but these preferences have a common pattern; namely, there is a most preferred alternative from which preference drops steadily in both directions. Put another way, of any three alternatives, the one in the intermediate position is never inferior to both of the others. Under this *single-peakedness* condition, majority voting defined a transitive relation and therefore an ordering. Hence, if the preferences of individuals are restricted to satisfy the single-peakedness condition, there does exist a constitution as defined earlier.

Statement of the Impossibility Theorem

I now state formally the conditions to be imposed on constitutions and then state the Impossibility Theorem, which simply asserts the non-existence of constitutions satisfying all of the conditions. The theorem as stated in the original paper (Arrow 1950) and in a subsequent book (Arrow 1951) is not correct as written, as shown by Blau (1957). To avoid confusion, I give a corrected statement and then explain the error.

Condition U: The constitution is defined for all logically possible profiles of preference orderings over the set of alternatives.

Condition M (Monotonicity): Suppose that x is socially preferred to y for a given profile. Now suppose a new profile in which x is raised in preference in some individual orderings and lowered in none. Then x is preferred to y in the social ordering associated with the new profile.

Condition I (Independence of Irrelevant Alternatives): Let S be a set of alternatives. Two profiles which have the same ordering of the alternatives in S for every individual determine the same social choice from S.

To state the next condition, it is necessary to define an imposed constitution as one in which there is some pair of alternatives for which the social choice is the same for all profiles.

Condition N (Non-imposition): The constitution is not imposed.

A constitution is said to be *dictatorial* if there is some individual, any one of whose strict preferences is the social preference according to that constitution.

Condition D (Non-dictatorship): The constitution is not dictatorial.

Theorem 1 There is no constitution satisfying Condition *U, M*, *I*, *N* and *C*.

A sketch of the argument can be given. From Condition *I,* the preference between any two alternatives depends only on the preferences of individuals between them and not on preferences about any other alternatives. Define a set of individuals to be *decisive* for alternative x against alternative y if the social preference is for x against y whenever all the individuals in the set prefer x to y. First, it can be shown that a set which is decisive for one alternative against one other is decisive for any alternative against any other. Hence, we can speak of a set of individuals as being decisive or not without reference to the alternatives being considered. If a set is not decisive, its complement (the voters not in the given set) can guarantee a weak preference, that is, preference or indifference. The set of all voters can easily be shown to be decisive, so there are decisive sets. The second stage in the proof is to take a decisive set with as few members of possible. If there were only one member, then by definition there would be a dictator, contrary to Condition *D*. Therefore, split the smallest decisive set so chosen into two subsets, say V_1 and V_2, and let V_3 contain all other voters. We now use an argument similar to that which showed the intransitivity of majority voting. Take any three alternatives, x, y and z. Suppose the members of V_1 all have the preference ordering, xyz, the members of V_2 the ordering yzx, and the members of V_3 the ordering zxy. Since V_1 and V_2 each have fewer members than the smallest

decisive set, neither is decisive. Since all voters other than those in V_2 prefer x to y, x must be preferred or indifferent socially to y. Since V_1 and V_2 together constitute a decisive set and y is preferred to z in both sets, y must be preferred socially to z. By transitivity, then, x is socially preferred to z. But x is preferred to z only by the members of V_1, which would therefore be decisive for x against z and hence a decisive set. This, however, contradicts the construction that V_1 is a proper subset of the smallest decisive set and therefore is not a decisive set. The theorem is therefore proved.

Notice that Condition *U,* that the constitution be defined for all profiles, is essential to the argument. We consider the consequences of particular profiles.

In Arrow (1951, p. 59), the theorem is stated with a weaker version of Condition *U* (and a corresponding restatement of Condition *M*).

Condition U′: The constitution is defined for a set of profiles such that, for some set of three alternatives, each individual can order the set in any way.

Since the contradiction requires only three alternatives, I supposed that the more general assumption would be sufficient. This is not so, as first pointed out by Blau (1957). The reason is that the non-dictatorship Condition *D* may hold for the set of all alternatives and not hold for a subset, such as the triple of alternatives just described. To illustrate, suppose there are four alternatives altogether. Let *S* be a set of three of them, and let *w* be the fourth. Suppose each individual may have any ordering such that *w* is either best or worst. There are two individuals in the society. The constitution provides that the social preference between any pair in *S* follows the preferences of individual 1, but *w* is best or worst according to individual 2's preference ordering. This constitution would satisfy all the conditions of the Arrow 1951 version and therefore provides a counter-ex. What is true, of course, is that individual 1 is a dictator over the alternatives in *S*. If we still wish to retain the weaker Condition *U′*, the theorem remains valid if a stronger non-dictatorship condition is imposed (see Murakami 1961).

Condition D′: No individual shall be a dictator over any three alternatives.

The conditions are fairly straightforward and need little comment. If it is reasonable to limit the range of possible individual orderings because of prior knowledge about the range of possible beliefs, then Condition *U* or *U′* could be replaced by a corresponding range condition. As has already been remarked, if preference orderings are restricted to the single-peaked type, then majority voting defines a constitution. There has been a considerable literature on range restrictions which imply that majority voting defines a constitution and some on more general voting methods. In a world of multi-dimensional issues, these restrictions are not particularly persuasive.

Conditions *M* and *N* embody different aspects of the value judgement that social decisions are made on behalf of the members of the society and should shift as values shift in a corresponding way. Condition *D* expresses a very minimal degree of democracy.

Condition *I* (independence of irrelevant alternatives) is central to the social choice approach whether in the Impossibility Theorem or in other, more positive, results. It is implicit in Rawls's difference principle of justice (Rawls 1971), as well as in utilitarianism or methods based on voting.

The above conditions have not included the Pareto principle explicitly.

Condition P: If every individual prefers x to y, then x is socially preferred to y.

It is not hard to prove, however, that this condition is implied by some of the previous conditions, specifically Conditions *M, I* and *N*. Further, if the Pareto condition is imposed, then the Impossibility Theorem holds without assuming Monotonicity or Non-imposition. Of course, it is obvious that the Pareto principle implies Non-imposition, since any choice can be enforced by unanimous agreement.

Theorem 2 There is no constitution satisfying Conditions *U, P, I, and D.*

This entry has dealt with Arrow's theorem itself and not with subsequent developments, which have been very abundant. The reader is referred to the entry on social choice in this work, and the surveys by Sen (1986) and Kelly (1978).

See Also

- ▶ Social Choice
- ▶ Social Welfare Function
- ▶ Welfare Economics

Bibliography

Arrow, K.J. 1950. A difficulty in the concept of social welfare. *Journal of Political Economy* 58: 328–346.

Arrow, K.J. 1951. *Social choice and individual values.* New York: Wiley. (2nd edn, New Haven: Yale University Press, 1963, identical except for an additional chapter.)

Arrow, K.J. 1973. Formal theories of social welfare. In *Dictionary of the history of ideas*, ed. P.P. Wiener, vol. 4, 276–284. New York: Scribner's.

Bergson (Burk), A. 1938. A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics* 52: 310–334.

Black, D. 1948. On the rationale of group decision-making. *Journal of Political Economy* 56: 23–34.

Black, D. 1958. *Theory of committees and elections*. Cambridge: Cambridge University Press.

Blau, J. 1957. The existence of social welfare functions. *Econometrica* 25: 302–313.

Bowen, H.R. 1943. The interpretation of voting in the allocation of economic resources. *Quarterly Journal of Economics* 58: 27–48.

Condorcet, M. de Caritat, Marquis de. 1785. *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix*. Paris: Imprimerie Royale.

Edgeworth, F.Y. 1897. The pure theory of taxation. In *Papers relating to political economy*, ed. F.Y. Edgeworth, vol. II, 63–125. London: Macmillan, 1925.

Hicks, J.R. 1939. The foundations of welfare economics. *Economic Journal* 49 (696–700): 711–712.

Kaldor, N. 1939. Welfare propositions of economics and interpersonal comparisons of utility. *Economic Journal* 49: 549–552.

Kelly, J.S. 1978. *Arrow impossibility theorems*. New York: Academic Press.

Murakami, Y. 1961. A note on the general possibility theorem of social welfare function. *Econometrica* 29: 244–246.

Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.

Scitovsky, T. 1941. A reconsideration of the theory of tariffs. *Review of Economic Studies* 9: 92–95.

Sen, A.K. 1986. Social choice theory. In *handbook of mathematical economics,* ed. K.J. Arrow and M. Intriligator, vol. III, 1073–81. Amsterdam, North-Holland.

Sidgwick, H. 1901. *Principles of political economy.* 3rd ed. London: Macmillan.

Arrow–Debreu Model of General Equilibrium

John Geanakoplos

Abstract

In the 1950s Kenneth Arrow and Gerard Debreu showed that the market system could be comprehensively analysed in terms of the neoclassical methodological premises of individual rationality, market clearing, and rational expectations, using the two mathematical techniques of convexity and fixed point theory. In so doing they greatly advanced the use of mathematics in economics.

Keywords

Agent optimization; Arrow–Debreu model of general equilibrium; Asymmetric information; Bankruptcy; Bounded rationality; Brouwer's fixed point theorem; Coalitions; Commodities; Comparative statics; Competitive equilibrium; Consumption loan model; Convexity; Core; Cournot's duopoly model; Differential pay principle; Existence of equilibrium; Externalities; Fixed point theorems; General equilibrium; Hicks, J. R.; Incomplete markets; Individualism; Insurance markets; Interpersonal utility comparisons; Kakutani's fixed point theorem; Lindahl equilibrium; Local uniqueness; Market clearing; Mathematical economics; Mathematics and economics; Microfoundations; Minkowski's theorem; Monotonicity; Negative prices; Neoclassical economics; Neoclassical production function; Non-satiation hypothesis; Ordinal utility; Overlapping generations model; Pareto optimality; Preference; Price; Rational expectations; Rational expectations equilibrium; Rationality; Risk allocation; Samuelson, P. A.; Sard's theorem; Separating hyperplane theorem; Tâtonnement; Transversality theory; Uncertainty; Utility; Walras's Law; Welfare economics; Welfare theorems

JEL Classifications
D5

Introduction

It is not easy to separate the significance and influence of the Arrow–Debreu model of general equilibrium from that of mathematical economics itself. In an extraordinary series of papers (Arrow 1951; Debreu 1951; Arrow and Debreu 1954), two of the oldest and most important questions of neoclassical economics, the viability and efficiency of the market system, were shown to be susceptible to analysis in a model completely faithful to the neoclassical methodological premises of individual rationality, market clearing, and rational expectations, through arguments at least as elegant as any in economic theory, using the two techniques (convexity and fixed point theory) that are still, after 30 years, the most important mathematical devices in mathematical economics. Fifteen years after its birth (for example, Arrow 1969), the model was still being reinterpreted to yield fresh economic insights, and 20 years later the same model was still capable of yielding new and fundamental mathematical properties (for example, Debreu 1970, 1974). When we consider that the same two men who derived the most fundamental properties of the model (along with McKenzie 1954) also provided the most significant economic interpretations, it is no wonder that its invention has helped earn for each of its creators, in different years, the Nobel Prize for economics.

In the next few pages I shall try to summarize the primitive mathematical concepts, and their economic interpretations, that define the model. I give a hint of the arguments used to establish the model's conclusions. Finally, on the theory that a model is equally well described by what it cannot explain, I list several phenomena that the model is not equipped to handle.

The Model

Commodities and Arrow–Debreu Commodities

(A.1) Let there be L commodities, $l = 1, \ldots, L$. The amount of a commodity is described by a real number. A list of quantities of all commodities is given by a vector in $\mathbb{R}^L$.

The notion of commodity is the fundamental primitive concept in economic theory. Each commodity is assumed to have an objective, quantifiable, and universally agreed upon (that is, measurable) description. Of course, in reality this description is somewhat ambiguous (should two apples of different sizes be considered two units of the same commodity, or two different commodities?) but the essential quantitative aspect of commodity cannot be doubted. Production and consumption are defined in terms of transformations of commodities that they cause. Conversely, the set of commodities is the minimum collection of objects necessary to describe production and consumption. Other objects, such as financial assets, may be traded, but they are not commodities. General equilibrium theory is concerned with the allocation of commodities (between nations, or individuals, across time, or under uncertainty, and so on). The Arrow–Debreu model studies those allocations which can be achieved through the exchange of commodities at one moment in time.

It is easy to see that it is often important to the agents in an economy to have precise physical descriptions of commodities, as for example when placing an order for a particular grade of steel or oil. The less crude the categorization of commodities becomes, the more scope there is for agents to trade, and the greater is the set of imaginable allocations. Two agents may each have apples and oranges. There is no point in exchanging one man's fruit for the other man's fruit, but both might be made better off if one could exchange his apples for the other's oranges. Of course there need not be any end to the distinctions which in principle could be drawn between commodities, but presumably finer details become less and less important. When the descriptions are so

precise that further refinements cannot yield imaginable allocations which increase the satisfaction of the agents in the economy, then the commodities are called Arrow–Debreu commodities.

A field is better allocated to one productive use than another depending upon how much rain has fallen on it; but it is also better allocated depending on how much rain has fallen on other fields. This illustrates the apparently paradoxical usefulness of including in the description of an Arrow–Debreu commodity characteristics of the world, for example the commodity's geographic location, its temporal location (Hicks 1939), its state of nature (Arrow 1953; Debreu 1959; Radner 1968), and perhaps even the name of its final consumer (Arrow 1969), which at first glance do not seem intrinsically connected with the object itself (but which are in principle observable).

Hicks, perhaps anticipated by Fisher and Hayek, was the first to suggest an elaborate notion of commodity; this idea has been developed by others, especially Arrow in connection with uncertainty. Hicks was also the first to understand apparently complicated transactions, perhaps involving the exchange of paper assets or other non-commodities, over many time periods, in terms of commodity trade at one moment in time. Thus saving, or the lending of money, might be thought of as the purchase today of a particular future dated commodity. The second welfare theorem, which we shall shortly discuss, shows that an 'optimal' series of transactions can always be so regarded. By making the distinction between the same physical object depending, for example, on the state of nature, the general equilibrium theory of the supply and demand of commodities at one moment in time can incorporate the analysis of the optimal allocation of risk (a concept which appears far removed from the mundane qualities of fresh fruit) with exactly the same apparatus used to analyse the exchange of apples and oranges. Classifying physical objects according to their location likewise allows transportation costs to be handled in the same framework. Distinguishing commodities by who ultimately consumes them could allow general equilibrium analysis to systematically include externalities and public goods as special cases, though this has not been much pursued.

In reality, it is very rare to find a market for a pure Arrow–Debreu commodity. The more finely the commodities are described, the less likely are the commodity markets to have many buyers and sellers (that is, to be competitive). More commonly, many groups of Arrow–Debreu commodities are traded together, in unbreakable bundles, at many moments in time, in 'second best' transactions. Nevertheless, this understanding of the limitations of real world markets, based on the concept of the Arrow–Debreu commodity, is one of the most powerful analytical tools of systematic accounting available to the general equilibrium theorist. Similarly, the model of Arrow–Debreu, with its idealization of a separate market for each Arrow–Debreu commodity, all simultaneously meeting, is the benchmark against which the real economy can be measured.

Consumers

(A.1) Let there be H consumers, $h = 1,\ldots, H$. Each consumer h can imagine consumption plans $x \in \mathbb{R}^L$ lying in some consumption set X^h.(A.2)X^h is a closed subset in $\mathbb{R}^L$ which is bounded from below.

Each consumer h also has well-defined preferences $\succcurlyeq h$ over every pair $(x,y) \in X^h \times X^h$, where $x \succcurlyeq y$ means x is at least as desirable as y. Typically it is assumed that (A.3) $\succcurlyeq$ is a complete, transitive, continuous ordering.

Notice that in general equilibrium consumers make choices between entire consumption plans, not between individual commodities. A single commodity has significance to the consumer only in relation to the other commodities he has consumed, or plans to consume. Together with transitivity and completeness, this hypothesis about consumer preferences embodies the neoclassical ideal of rational choice.

Rationality has not always been a primitive hypothesis in neoclassical economics. It was customary (for example, for Bentham, Jevons, Menger, Walras) to regard satisfaction, or utility,

as a measurable primitive; rational choice, when it was thought to occur at all, was the consequence of the maximization of utility. And since utility was often thought to be instantaneously produced, sequential consumer choice on the basis of sequential instantaneous utility maximization was sometimes explicitly discussed as irrational (see, for example, Böhm-Bawerk on saving and the reasons why the rate of interest is always positive).

Once utility is taken to be a function not of instantaneous consumption, but of the entire consumption plan, then rational choice is equivalent to utility maximization. Debreu (1951) proved that any preference ordering $\succcurlyeq_h$ defined on $X^h \times X^h$ satisfies (A.1)–(A.3) if and only if there is a utility function $u^h : X^h \to \mathbb{R}$ such that $x \succcurlyeq_h y$ exactly when $u^h(x) \geqq u^h(y)$.

Under the influence of Pareto (1909), Hicks (1939) and Samuelson (1947), neoclassical economics has come to take rationality as primitive, and utility maximization as a logical consequence. This has had a profound effect on welfare economics, and perhaps on the scope of economic theory as well. In the first place, if utility is not directly measurable, then it can only be deduced from observable choices, as in the proof of Debreu. But at best this will give an 'ordinal' utility, since if $f: \mathbb{R} \to \mathbb{R}$ is any strictly increasing function, then u^h represents $\succcurlyeq_h$ if and only if $v^h \equiv f^o u^h$ represents $\succcurlyeq_h$. Hence there can be no meaning to interpersonal utility comparisons; the Benthamite sum $\Sigma_{h=1}^{H} u^h$ is very different from the Benthamite sum $\Sigma_{h=1}^{H} f^{ho} u^h$. In the second place, the ideal of rational choice or preference, freed from the need for measurement, is much more easily extended to domains not directly connected to the market and commodities such as political candidates or platforms, or 'social states'. The elaboration of the nature of the primitive concepts of commodity and rational choice, developed as the basis of the theory of market equilibrium, prepared the way for the methodological principles of neoclassical economics (rational choice and equilibrium) to be applied to questions far beyond those of the market.

Although the rationality principle is in some respects a weakening of the hypothesis of measurable utility and instantaneous utility maximization, when coupled with the notion of consumption plan it is also a strengthening of this hypothesis, and a very strong assumption indeed. For example there is not room in this theory for the Freudian split psyche (or self-deception), or for Odysseus-like changes of heart. Perhaps more importantly, a consumer's preferences (for example how thrifty he is) do not change according to the role he plays in the process of production (for example, on whether he is a capitalist or landowner), nor do they change depending on other consumers' preferences, or the supply of commodities. As an instance of this last case, note that it follows from the rationality hypothesis that the surge in the microcomputer industry influenced consumer choice between typewriters and word processors only through availability (via the price), and not through any learning effect. (Consumers can 'learn' in the Arrow–Debreu model, for example their marginal rates of substitution can depend on the state of nature, but the rate at which they learn is independent of production or consumption – it depends on the exogenous realization of the state. We shall come back to this when we consider information.) If for no other reason, the burden of calculation and attention which rational choice over consumption plans imposes on the individual is so large that one expects rationality to give way to some kind of bounded rationality in some future general equilibrium models.

Two more assumptions on preferences made in the model of Arrow–Debreu are nonsatiation and convexity:

(A.4) For each $x \in X^h$, there is a $y \in X^h$ with $y \succ_h x$, that is, such that $y \succcurlyeq_h x$ and not $x \succcurlyeq_h y$

(A.5) X^h is a convex set, and $\succcurlyeq_h$ is convex, that is, if $y \succ_h x$ and $0 < t \leq 1$, then $[ty + (1-t)x] \succ_h x$.

The nonsatiation hypothesis seems entirely in accordance with human nature. The convexity hypothesis implies that commodities are infinitely divisible, and that mixtures are at least as good as extremes. When commodities are distinguished very finely according to dates, so that they must be thought of as flows, then the convexity hypothesis is untenable. In a standard example, a man

may be indifferent between drinking a glass of gin or of scotch at a particular moment, but he would be much worse off if he had to drink a glass of half gin–half scotch. On the other hand, if the commodities were not so finely dated, then they would be more analogous to stocks, and a consumer might well be better off with a litre of gin and a litre of scotch, than two litres of either one. In any case, as we shall remark later, if every agent is small relative to the market (that is, if there are many agents) then the non-convexities in preferences are relatively unimportant.

Each agent h is also characterized by a vector of initial endowments

(A.6) $e^h \in X^h \subset \mathbb{R}^L$ for all $h = 1,\ldots, h$.

The endowment vector e^h represents the claims that the consumer has on all commodities, not necessarily commodities in his physical possession. The fact that $e^h \in X^h$ means that the consumer can ensure his own survival even if he is deprived of all opportunity to trade. This is a somewhat strange hypothesis for the modern world, in which individuals often have labour but few other endowments, e.g. land. Doubtless the hypothesis could be relaxed; in any case, survival is not an issue that is addressed in the Arrow–Debreu model.

Each individual h is also endowed with an ownership share of each of the firms $j = 1, \ldots, J$

(A.7) For all $h = 1,\ldots, H, j = 1,\ldots, J$, $d_{hj} \geq 0$, and for all $j = 1,\ldots, J$, $\Sigma_{h=1}^{H} d_{hj} = 1$.

Firms. (A.8) Let there be J firms, $j = 1,\ldots, J$.

The firm in Arrow–Debreu is characterized by its initial distribution of owners, and by its technological capacity $Y_j \subset \mathbb{R}^L$ to transform commodities. Any production plan $y \in \mathbb{R}^L$, where negative components of y refer to inputs and positive components denote outputs, is feasible for firm j if $y \in Y_j$. A customary assumption made in the Arrow–Debreu model is free disposal: if $l = 1,\ldots, L$ is any commodity, and v_1 is the unit vector in $\mathbb{R}^L$, with one in the lth coordinate and zero elsewhere, then

(A.9) For all $l = 1,\ldots, L$ and $k > 0$, $-kv_l \in Y_j$, for some $j = 1,\ldots, J$.

Although it is strange, when thinking of nuclear waste etc., to think that any commodity can be disposed without cost (i.e. without the use of any other inputs), as we shall remark later, this assumption can be relaxed, if negative prices are introduced (or if weak monotonicity is assumed).

The empirically most vulnerable assumption to the Arrow–Debreu model, and one crucial to its logic, is:

(A.10) For each j, Y_j is a closed, convex set containing 0.

This convexity assumption rules out indivisibilities in production (e.g. half a tunnel), increasing returns to scale, gains from specialization, etc. As with consumption, if the indivisibilities of production are small relative to the size of the whole economy, then the conclusions we shall shortly present are not much affected. But when they are large, or when there are significant increasing returns to scale, the model of competitive equilibrium that we are about to examine is simply not applicable. Nevertheless, convexity is consistent with the traditionally important cases of decreasing and constant returns to scale in production.

We conclude by presenting three final assumptions used in the Arrow–Debreu model.

(A.11) Let $e = \Sigma_{h=1}^{H} e^h$,
let $F = \left\{ y \in \mathbb{R}^L \middle| \, y = \Sigma_{j=1}^{J} y_j, y_j \in Y^j, j = 1,\ldots,J \right\}$,
let $\overline{F} = \{ y \in F \,|\, y + e \geqq 0 \}$, and
let $K = \left\{ (y_1, \ldots, y_J) \in Y_1 \times \ldots \times Y_J \middle| \, \equiv y \sum_{j=1}^{J} y^j \in \overline{F} \right\}$.
Then $\overline{F} \cap \mathbb{R}^L_{++} \neq \phi$, and K is compact.

Debreu model.

Assumption (A.11) requires that the level of productive activity that is possible even if the productive sector appropriates all the resources of the consuming sector is bounded (as well as closed).

Notice that these assumptions are consistent with firms owning initial resources, as well as individuals. In the original Arrow–Debreu model

(1954), the firms were prohibited from owning initial resources (they were assigned to the firm owners: with complete markets there is little difference, but with incomplete markets the earlier assumption is restrictive).

(A.12) The economy is irreducible.

We shall not elaborate this assumption here. It means that for any two agents h and h', the endowment e^h of agent h is positive in some commodity l, which (taking into account the possibilities of production) agent h' could use to make himself strictly better off. It certainly seems reasonable that each agent's labour power could be used to make another agent better off.

Lastly, we assume that

(A.13) The commodities are not distinguished according to which firm produces them, or who consumes them.

Assumption (A.13) is made simply for the purposes of interpretation. When put together with the definition of competitive equilibrium, it implies that there are no externalities to production or consumption, no public goods, etc. Mathematically, however, (A.13) has no content. In other words, if we dropped assumption (A.13), the Arrow–Debreu notion of competitive equilibrium would still make sense (even in the presence of externalities and public goods) and it would still have the optimality properties we shall elaborate in section "Equilibrium", but it would require an entirely different interpretation. Consumers, for example, would be charged different prices for the same physical commodities (same, that is, according, to date, location and state of nature). In more technical language, a Lindahl equilibrium is a special case of an (A.1)–(A.12) Arrow–Debreu equilibrium, with the commodity space suitably expanded and interpreted. Thus each physical unit of a public good is replaced by H goods, one unit for the public good indexed by which agent consumes it. Also the physical technology set describing the production of the public good is replaced by a different set in the Arrow–Debreu model, lying in a higher dimensional space, where the output of the one physical public good is replaced by the joint output of the same amount of – goods. In an Arrow–Debreu equilibrium, consumers will likely pay different prices for these H goods, i.e. for what in reality represents the same physical public good. Hence the differential pay principle for the optimal provision of public goods elucidated by Samuelson, which appeared to point to a qualitative difference between the analytical apparatus needed to describe optimality in public goods and private goods economies, is thus shown to be explicable by exactly the same apparatus used for private goods economies, simply by multiplying the number of commodities. The same device can also be used for analysing the optimal provision of goods when there are externalities, provided that negative prices are allowed. Assumption (A.13) thus seriously limits the normative conclusions that can be drawn from the model. From a descriptive point of view, however, rationality and the price taking behaviour which equilibrium implies make (A.13) necessary.

Equilibrium

Price is the final primitive concept in the Arrow–Debreu model. Like commodity it is quantifiable and directly measurable. As Debreu has remarked, the fundamental role which mathematics plays in economics is partly owing to the quantifiable nature of these two primitive concepts, and to the rich mathematical relationship of dual vector spaces, into which it is natural to classify the collections of price values and commodity quantities. Properly speaking, price is only sensible (and measurable) as a relationship between two commodities, i.e. as relative price. Hence there should be $L^2 - L$ relative prices in the Arrow–Debreu model. But the definition of Arrow–Debreu equilibrium immediately implies that it suffices to give $L - 1$ of these ratios, and all the rest are determined.

For mathematical convenience (namely to treat prices and quantities as dual vectors), one price is specified for each unit quantity of each commodity. The relative price of two commodities can be

obtained by taking the ratio of the Arrow–Debreu prices of these commodities. I shall proceed by specifying the definition of Arrow–Debreu equilibrium, and then I make a number of remarks emphasizing some of the salient characteristics of the definition. The longest remark concerns the differences between the historical development of general equilibrium, up until the time of Hicks and Samuelson and the particular Arrow–Debreu model of general equilibrium.

An Arrow–Debreu economy E is an array $E = \{L, H, J\ (X^h, e^h, \succ_h), (Y^j), (d^{hj}), h = 1,H, j = 1,\ldots,J\}$ satisfying assumptions (A.1)–(A-13). An Arrow–Debreu equilibrium is an array $\left[(\overline{p}_l), (\overline{x}_l^h), \left(\overline{y}_l^j\right), l = 1, \ldots, L, h = 1, \ldots, H, j = 1, \ldots, J\right]$ satisfying:

For all $j = 1, ..., J, \overline{y}_j \in$

$$\arg \max \left\{ \sum_{l=1}^{L} \overline{p}_l y_1 \,|\, (y = y_1, \ldots, y_L) \in Y^j \right\} \quad (1)$$

For all $h = 1, \ldots, H \overline{x}^h \in B^h(\overline{p})$, where $B^h(\overline{p})$

$$\equiv \left\{ x \in X^h \,|\, \sum_{l=1}^{L} \overline{p}_l x_l \leqq \sum_{l=1}^{L} \overline{p}_l e_l^h + \sum_{j=1}^{J} d^{hj} \sum_{l=1}^{L} \overline{p}_1 \overline{y}_l^j \right\} \quad (2)$$

and if $x \in B^h(\overline{p})$, then not $x \succ h\overline{x}^h$,

$$\text{For all } l = 1, ..., L, \sum_{h=1}^{H} \overline{x}_l^h = \sum_{h=1}^{H} e_l^h + \sum_{j=1}^{J} y_l^{-j}. \quad (3)$$

The most striking feature of general equilibrium is the juxtaposition of the great diversity in goals and resources it allows, together with the supreme coordination it requires. Every desire of each consumer, no matter how whimsical, is met precisely by the voluntary supply of some producer. And this is true for all markets and consumers simultaneously.

There is a symmetry to the general equilibrium model, in the way that all agents enter the model individually motivated by self-interest (not as members of distinct classes motivated by class interests), and simultaneously, so that no agent acts prior to any other on a given market (e.g. by setting prices). If workers' subsistence were not assumed, for example, that would break the symmetry; workers income could have to be guaranteed first, otherwise demand would (discontinuously) collapse. As it is, at the aggregate level, supply and demand equally and simultaneously determine price; in equilibrium, both the consumers' marginal rates of substitution and the producers' marginal rates of transformation are equal to relative prices (assuming differentiability and interiority). There are gains to trade both through exchange and through production. This point of view represents a significant break with the classical tradition of Ricardo and Marx. We shall come to the main difference between the classical and neoclassical approaches shortly. Another difference is that there need not be fixed coefficients of production in the Arrow–Debreu model – the sets Y are much more general. Also in an Arrow–Debreu equilibrium, there is no reason for there to be a uniform rate of profit. There is none the less one aspect of the model which these authors would have greatly approved, namely the shares d^{hj} which allow the owners of firms to collect profits even though they have contributed nothing to production.

Notice that in general equilibrium each agent need only concern himself with his own goals (preferences or profits) and the prices. The implicit assumption that every agent 'knows' all the prices is highly non-trivial. It means that at each date each agent is capable of forecasting perfectly all future prices until the end of time. It is in this sense that the Arrow–Debreu model depends on 'rational expectations'. Each agent must also be informed of the 'price qj of each firm j, where $q_j = \Sigma_{l-1}^{L} \overline{p}_l \overline{y}_l^j$. (Firms that produce under constant returns to scale must also discover the level of production, which cannot be deduced from the prices alone.) Assuming that the 'man on the spot' (Hayek's expression) knows much better than anyone else what he wants, or best how his changing environment is suited to producing his product, decentralized decision making would seem to be highly desirable, if it is not incompatible with coordination. Indeed, harmony through

diversity is one of the sacred doctrines of the liberal tradition.

The greatest triumph of the Arrow–Debreu model was to lay out explicitly the conditions (roughly (A.1)–(A.13)) under which if is possible to claim that a properly chosen price system must always exist that, like the invisible hand, can guide diverse and independent agents to make mutually compatible choices. The idea of general equilibrium had gradually developed since the time of Adam Smith, mostly through the pioneering work of Walras (1874), von Neumann (1937), Hicks (1939) and Samuelson (1947). By the late 1940s the definition of equilibrium, including ownership shares in the firms, was well-established. But it was Arrow–Debreu (1954) that spelled out precise microeconomic assumptions at the level of the individual agents that could be used to show the model was consistent.

The axiomatic and rigorous approach that characterized the formulation of general equilibrium by Arrow–Debreu has been enormously influential. It is now taken for granted that a model is not properly defined unless it has been proved to be logically consistent. Much of the clamour for 'microeconomic foundations to macroeconomics', for example, is a desire to see an axiomatic clarity similar to that of the Arrow–Debreu model applied to other areas of economics. Of course, there were other earlier economic models that were similarly axiomatic and rigorous; one thinks especially of von Neumann–Morgenstern's *Theory of Games* (1944). But game theory was, at the time, on the periphery of economics. Competitive equilibrium is at its heart.

The central mathematical techniques, convexity theory (separating hyperplane theorem) and Brouwer's (Kakutani's) fixed point theorem, used in Arrow–Debreu are, 30 years later, still the most important tools used in mathematical economics. Both elements had played a (hidden) role in von Neumann's work. Convexity had been prominent in the work of Koopmans (1951) on activity analysis, in the work of Kuhn and Tucker (1951) on optimization, and in the papers of Arrow (1951) and Debreu (1951) on optimality. Fixed point theorems had been used by von Neumann (1937), by Nash (1950) and especially by McKenzie (1954), who one month earlier than Arrow–Debreu had published a proof of general equilibrium using Kakutani's theorem, albeit in a model where the primitive assumptions were made on demand functions, rather than preferences. McKenzie (1959) also made an early contribution to the notion of an irreducible economy (assumption (A.9)).

The first fruit of the more precise formulation of equilibrium that began to emerge in the early 1950s was the transparent demonstration of the first and second welfare theorems that Arrow and Debreu simultaneously gave in 1951. Particularly noteworthy is the proof that every equilibrium is Pareto optimal. So simple and illuminating is this demonstration that it is no exaggeration to call it the most frequently imitated argument in all of neoclassical economic theory.

Among the confusions that were cleared away by the careful axiomatic treatment of equilibrium was the reliance of the discussions by Hicks and Samuelson on interior solutions and differentiability. When discussing the optimal allocation of housing, for example, it is evident that most agents will consume nothing of most houses, but this does not affect the Pareto optimality of a free (and complete) market allocation of housing. Similarly, it is not necessary to either the existence of Arrow–Debreu equilibrium, nor to the first and second welfare theorems, that preferences or production sets be either differentiable or strictly convex. In particular, it is possible to incorporate the 'neoclassical production function' with constant returns to scale with variable inputs, the classical fixed coefficients methods of production, and the strictly concave production functions of the Hicks-Samuelson vintage, all in the same framework.

This is not to say that differentiability has no role to play in the Arrow–Debreu model. In his seminal paper (1970), Debreu resurrected the role of differentiability by showing, via the methods of transversality theory (a branch of differential topology) that almost every differentiable economy is regular, in the sense that small perturbations to the economic data (e.g. the endowments)

make small changes in all the equilibrium prices. Before Debreu, comparative statics could be handled only under specialized hypotheses, for example, the invertibility of excess demand at all prices, etc. We shall give a fuller discussion of the three crucial mathematical results of the Arrow–Debreu model – existence, optimality and local uniqueness – in the next section.

Observe finally, that although the commodities may include physical goods dated over many time periods, there is only one budget constraint in an Arrow–Debreu equilibrium. The income that could be obtained from the sale of an endowed commodity, dated from the last period, is available already in the first period.

Pareto Optimality

The first theorem of welfare economics states that any Arrow–Debreu equilibrium allocation $\bar{x} = (\bar{x}^h)$, $h = 1,\ldots, H$ is Pareto optimal in the sense that if $[(x^h), (y^j)]$ satisfies $y^j \in Y^j, \Sigma_{h=1}^{H} x^h = \Sigma y_n^j + e$, then it cannot be the case that $x^h \succ h^{\bar{x}^h}$ for all h. The second theorem of welfare analysis states the converse, namely that any Pareto optimal allocation for an Arrow–Debreu economy E is a competitive equilibrium allocation for an Arrow–Debreu economy $\hat{E}$ obtained from E by rearranging the initial endowments of commodities and ownership shares.

The first welfare theorem expresses the efficiency of the ideal market system, although it makes no claim as to the justice of the initial distribution of resources. The second welfare theorem implies that any income redistribution is best effected through a lump sum transfer, rather than through manipulating the market, e.g. through rent control, etc.

The connection between competitive equilibrium and Pareto optimality has been perceived for a long time, but until 1951 there was a general confusion between the necessity and sufficiency part of the arguments. The old proof of Pareto optimality (see Lange 1942) assumed differentiable utilities of production sets, and a strictly positive allocation $\bar{x}$. It noted the first order conditions to the problem of maximizing the ith consumer's utility, subject to maintaining all the others at least as high as they got under $\bar{x}$, and feasibility, are satisfied at $\bar{x}$, if and only if $\bar{x}$; is a competitive equilibrium allocation for a 'rearranged' economy $\hat{E}$. This first order, or infinitesimal, proof of equivalence between competitive equilibrium and Pareto optimality could have been made global by postulating in addition that preferences and production sets are convex.

The Arrow and Debreu (1954) proofs of the equivalence between competitive equilibrium and Pareto optimality, under global changes, do not require differentiability, nor do they require that all agents consume a strictly positive amount of every good. In fact the proof of the first welfare theorem, that each competitive equilibrium is Pareto optimal, does not even use convexity.

The only requirement is local nonsatiation, so that every agent spends all his income in equilibrium. If (x, y) Pareto dominates the equilibrium allocation $(\bar{p},\bar{x},\bar{y})$, then for all $h, \bar{p} \cdot x^h < \bar{p} \cdot \bar{x}^h$. Since profit maximization implies that for all $j, \bar{p} \cdot \bar{y}^j \geqq \bar{p} \cdot y^j$, it follows that $\bar{p} \cdot \left(\Sigma_h x^h - \Sigma_j y^j\right) > \bar{p} \cdot \left(\Sigma_h \bar{x}^h - \Sigma_h \bar{y}^j\right)$ contradicting feasibility.

The proof of the second welfare theorem, on the other hand, does require convexity of the preferences and production sets (though not their differentiability, nor the interiority of the candidate allocation $\bar{x}$). Essentially it depends on Minkowski's theorem, which asserts that between any two disjoint convex sets in $\mathbb{R}^L$ there must be a separating hyperplane.

In this connection let us mention one more remarkable mathematical property of the Arrow–Debreu model. Let us suppose that all production takes place under constant returns to scale: if $y \in Y^j$, then so is λy, for $\lambda \geq 0$. We say that a feasible allocation $\bar{x}$ for the economy E is in the core if there is no coalition of consumers $S \subset \{1,\ldots, H\}$ such that using only their initial endowments of resources, as well as access to all the production technologies, they cannot achieve an allocation for themselves which they all prefer to $\bar{x}$. The core is meant to reflect those allocations which could be maintained when bargaining (the formation of coalitions) is costless. In a status quo core allocation,

any labour union or cartel of owners that threatens to withhold its goods from the market knows that another coalition could form and by withholding its goods, prevent some members of the original coalition from being better off than they were under the status quo. It is easy to see that any competitive equilibrium is in the core. Debreu–Scarf (1963), building on earlier work of Scarf, showed by using the separating hyperplane theorem, that if agents are small relative to the market, in the sense they made precise through the notion of replication, then the core consists only of competitive allocations. Such a theorem can also be proved even if there are small nonconvexities in preferences (see Aumann 1964, for a different formulation of the small agent).

Existence of Equilibrium

Suppose that agents' preferences and firms' production sets are strictly convex, and that agents strictly prefer more of any commodity to less (strict monotonicity) and that they all have strictly positive endowments. Let Δ be the set of L-price vectors, all non-negative, summing to one. Let $f^h(p)$ be the commodity bundle most preferred by agent h, given the strictly positive prices $p \in \Delta_{++}$. Similarly let $g^j(p)$ be the profit maximizing choice of firm j, given prices $p \in \Delta_{++}$. Finally, let $f(p) = \Sigma_{h=1}^{H} f^h(p) - \Sigma_{j=1}^{j} g^j(p) - e$. It is easy to show that f is a continuous function at all $p \in \Delta_{++}$. A price $\overline{p} \in \Delta_{++}$ is an Arrow–Debreu equilibrium price if and only if $f(\overline{p}) = 0$.

In general there is no reason to expect a continuous function to have a zero. Thus Wald could prove only with great difficulty in a special case that an equilibrium necessarily exists. Now observe that the function must satisfy Walras's Law, $p \cdot f(p) = 0$, for all p. So f is not arbitrary.

Consider the convex, compact set Δe of prices $p \in \Delta$ with $p_l \geqq \varepsilon > 0$, for all l. Consider also the continuous function $\phi: \Delta_e \to \Delta_e$ mapping p to the closest point $\hat{p}$ in Δe to $f(p) + p$. By Brouwer's fixed point theorem, there must be some $\overline{p}$ with $\phi(\overline{p}) = \overline{p}$. From strict monotonicity, it follows that $\overline{p}$ cannot be on the boundary of Δe, if ε is chosen sufficiently small. From Walras's Law it follows that if $\overline{p}$ is in the interior of Δe, then $f(\overline{p}) = 0$: The demonstration of the existence of equilibrium by Arrow and Debreu, as modified later by Debreu (1959), followed a similar logic.

Note the essential role of convexity in two parts of the above proof. It was used with respect to agents' characteristics to guarantee that their optimizing behaviour is continuous. And it was also used to ensure that the space Δ has the fixed point property. Smale (1976) has given a path-following proof (related to Scarf's, 1973, algorithm) that on closer inspection does not require convexity of the price space. (Dierker 1974; and Balasko 1986, have given homotopy proofs.) This is not only of computational importance. It appears that there may be economic problems, dealing with general equilibrium with incomplete markets, in which the price space is intrinsically nonconvex, and in which the existence of equilibrium can only be proved using path-following methods (see Duffie–Shaffer 1985).

To weaken the assumption of strict convexity, in the above proof, one can replace Brouwer's fixed point theorem with Kakutani's. An important conceptualpoint arises in connection with strict monotonicity. If that is dropped, and the production sets do not have free disposal, then, in order to guarantee the existence of equilibrium, the definition must be revised to require either $f_l(\overline{p}) = 0$, or $f_l(\overline{p}) < 0$ and $\overline{p}_l = 0$. There may be free goods, like air, in excess supply. One cannot drop monotonicity and free disposal without allowing for negative prices.

Finally, it can be shown that if there are small nonconvexities in either preference or production, and if all the agents are small relative to the market (either in the replication sense of Debreu–Scarf, or the measure zero sense of Aumann), then there will be prices at which the markets nearly clear. On the other hand, increasing returns to scale over a broad range is definitely incompatible with equilibrium.

Local Uniqueness and Comparative Statics

Another property of the excess demand function $f(p)$ is that it is homogeneous of degree zero. So instead of taking $p \in \Delta$, let us fix $p_1 = 1$. Similarly, let $F(p)$ be the $L - 1$ vector of excess

demands for goods $l = 2, \ldots, L$. If $F(p) = 0$, then by Walras's Law, $f(p) = 0$.

Suppose furthermore that agent characteristics are smooth. Then $F(P)$ is a differentiable function. If $D_pF(\overline{p})$ has full rank at an equilibrium $\overline{p}$, then $\overline{p}$ is locally unique. Moreover, the equilibrium $\overline{p}$ will move continuously, given continuous, small changes in the agents' characteristics, such as their endowments e. If $D_pF(\overline{p})$ has full rank at all equilibria $\overline{p}$, then there are only a finite number of equilibria. Debreu (1970) called an economy E regular if $D_pF(\overline{p})$ has full rank at all equilibrium $\overline{p}$ of E.

The problem of trying to give sufficient conditions on preferences etc. to guarantee that $D_p F$ has full rank in equilibrium has proved intractable (except for restrictive, special cases). But Debreu (1970) solved the problem in classic style, appealing to the transversality theorem of differential topology (or Sard's theorem), to show that if one were content with regularity for 'almost all' economies, then the problem is simple. He proved that for almost all economies, $D_p F$ has full rank at every equilibrium. Hence, in almost all economies comparative statics (the change in equilibrium, given exogenous changes to the economy) is well defined.

Observe that excess demand F depends on the agents' characteristics, including their endowments, so we could write $F(e, p)$. Now the transversality theorem says that (given some technical conditions) if $D_eF(e, \overline{p})$ has full rank at all equilibria $\overline{p}$ for the economy $E(e)$ with endowments e, for all e, then for 'almost all' e, $D_eF(e,\overline{p})$ has full rank at all equilibrium $\overline{p}$ for $E(e)$. But it is easy to show that $D_eF(e,\overline{p})$ always has full rank. Along similar lines, Debreu proved that the 'generic regularity' of equilibrium.

There is one unfortunate side to this comparative statics story. One would like to show not only that comparative statics are well defined, but also that they have a definite form. In a concave programming problem, for example, a small increase in an input results in a decrease in that input's shadow price, and an increase in output approximately equal to the size of the input increase multiplied by its original shadow price. Given the strong rationality hypothesis of the Arrow–Debreu model, one would hope for some sort of analogous result. Following a conjecture of Sonnenschein, Debreu proved in 1974 that given any function $f(p)$ on Δe, satisfying Walras's Law, he could find an Arrow–Debreu economy such that $f(p)$ is its aggregate excess demand on Δe. This assumptions (A.1)–(A.13) do not permit any a priori predictions about the changes that must occur in equilibrium given exogenous changes to the economy. An increase in the aggregate endowment of a particular good, for example, might cause its equilibrium price to rise. The possibility of such pathologies is disappointing. It means that to make even qualitative predictions, the economist needs detailed data on the excess demands F.

What the Model Doesn't Explain

We have already discussed the implications of the notion of Arrow–Debreu commodities and the second welfare theorem for insurance, namely that since every Pareto optimal allocation is supportable as an Arrow–Debreu equilibrium, every optimal allocation of risk bearing can be accomplished by the production and trade of Arrow–Debreu commodities, i.e. without recourse to additional kinds of insurance markets specializing in risks. Every Arrow–Debreu commodity is as much a diversifier in location, or time, or physical quality as it is for risk. This leads to a great simplification and economy of analysis. But it also means that, from the positive point of view, the Arrow–Debreu economy cannot directly provide an analysis of insurance markets (except as a benchmark case). In this section "Introduction" shall try to point out a few of the other phenomena which needle into the background in the Arrow–Debreu model, but which would emerge if the assumption of a finite, but complete set of Arrow–Debreu commodities, and consumers was dropped.

There are four currently active lines of research which attempt to come to grips in a general equilibrium framework with some of these phenomena, while preserving the fundamental neoclassical Arrow–Debreu principles of agent optimization,

market clearing, and rational expectations, that I think are particularly worthy of attention. They are the theory of general equilibrium with incomplete asset markets which can be traced back to Arrow's (1953) seminal paper on securities; overlapping generations economies, whose study was initiated by Samuelson (1958) in his classic consumption loan model; the Cournot theory of market exchange with few traders, first adapted to general equilibrium by Shapley–Shubik (1977), and the model of rational expectations equilibrium, pioneered by Lucas (1972).

Let us note first of all that in Arrow–Debreu equilibrium there is no trade in shares of firms. A stock certificate is not an Arrow–Debreu commodity, for its possession entitles the owner to additional commodities which he need not obtain through exchange. Note also that in Arrow–Debreu equilibrium, the hypothesis that all prices will remain the same, no matter how an individual firm changes its production plan, guarantees that firm owners unanimously agree on the firm objective, to maximize profit. If there were a market for firm shares, there would not be any trade anyway, since ownership of the firm and the income necessary to purchase it would be perfect substitutes. In an incomplete markets equilibrium, different sources of revenue are not necessarily perfect substitutes. There could be active trade on the stock market. Of course, such a model would have to specify the firm objectives, since one would not expect unanimity. The theory of stock market equilibrium is still in its infancy, although some important work has already been done (See Drèze 1974; and Grossman–Hart 1979).

Bankruptcy is not allowed in an Arrow–Debreu equilibrium. That follows from the fact that all agents must meet their budget constraints. In a game theoretic formulation of equilibrium (such as I shall discuss shortly), it is achieved by imposing an infinite bankruptcy penalty. Since every Arrow–Debreu equilibrium is Pareto optimal, there would be no benefit in reducing the bankruptcy penalty to the point where someone might choose to go bankrupt. But with incomplete markets, such a policy might be Pareto improving, even allowing for the deadweight loss of imposing, the penalties.

Money does not appear in the Arrow–Debreu model. Of course, all of the reasons for its life existence: transactions demand, precautionary demand, store of value, unit of account, etc. are already taken care of in the Arrow–Debreu model. One could imagine money in the model: at data zero every agent could borrow money from the central bank. At every date afterwards he would be required to finance his purchases out of his stock of money, adding to that stock from his sales. At the last data he would be required to return to the bank exactly what he borrowed (or else face an infinite bankruptcy penalty). In such a model the Arrow–Debreu prices would appear as money prices. The absolute level of money prices and the aggregate amount of borrowing would not be determined, but the allocations of commodities would be the same as in Arrow–Debreu. There is no point in making the role of money explicit in the Arrow–Debreu model, since it has no effect on the real allocations. However, if one considers the same model with incomplete asset markets, the presence of explicitly financial securities can be of great significance to the real allocations.

In the Arrow–Debreu model, all trade takes place at the beginning of time. If markets were reopened at later dates for the same Arrow–Debreu commodities, then no additional trade would take place anyway. At the other extreme, one might consider a model in which at every date and state of nature only those Arrow–Debreu commodities could be traded which were indexed by the corresponding (date, state) pair. An intermediate case would also permit the trade of some (but not all) differently indexed Arrow–Debreu commodities. Now the Arrow–Debreu proofs of the existence and Pareto optimality of equilibrium do not apply to such an incomplete markets economy, as Hart (1975) first pointed out. We have already noted the existence problem. As for efficiency, the Pareto optimality of Arrow–Debreu equilibria might suggest the presumption that, though there might be a loss to eliminating markets, trade on the remaining markets would be as efficient as possible. In fact, it can be shown (generically) that equilibrium trade do not make efficient use of the existing markets.

The Arrow–Debreu model of general equilibrium is relentlessly neoclassical; in fact it has become the paradigm of the neoclassical approach. This stems in part from its individualistic hypothesis, and its celebrated conclusions about the potential efficacy of unencumbered markets (Although Arrow, for example, has always maintained that a proper understanding of Arrow–Debreu commodities is also useful to showing how inefficient is the limited real world market system). But still more telling is the fact that the assumption of a finite number of commodities (and hence of dates) forces upon the model the interpretation of the economic process as a one- way activity of converting given primary resources into final consumption goods. If there is universal agreement about when the world will end, there can be no question about the reproduction of the capital stock. In equilibrium it will be run down to zero. Similarly when the world has a definite beginning, so that the first market transaction takes place after the ownership of all resources and techniques of production, and the preferences of all individuals have been determined, one cannot study the evolution of the social norms of consumption in terms of the historical development of the relations of production. One certainly cannot speak about the production of all commodities by commodities (Sraffa 1960) (since at date zero there must be commodities which have not been produced by commodities, i.e. by physical objects which are traded).

It seems natural to suppose that as L becomes very large, so that the end of the world is put off until the distant future, that this event cannot be of much significance to behaviour now. But let us not forget the rationality imposed on the agents. Far off as the end of the world might be, it is perfectly taken into account. Thus, for example, social security (funded as it is in the US by taxes on the young) could not exist if rational agents agreed on a final stopping time to transactions.

Consider a model satisfying all the assumptions (A.1)–(A.13), except that L and H are allowed to be infinite, such as the overlapping generations model. It can be shown that there is a robust collection of economies which have a continuum of equilibria, most of which are Pareto sub-optimal, which differ enormously in time 0 behaviour. Thus in a model where time does not have a definite end, the optimality and comparative statics properties of equilibria are radically different (For example, there may be a continuum of equilibria, indexed by the level of period 0 real wages – inversely related to the rate of profit – or the level of output or employment. The interested reader can consult the entry on the overlapping generations model of general equilibrium. A systematic study of economies where only L is allowed to be infinite was begun by Bewley (1972). Such economies tend to have properties similar to those of Arrow–Debreu).

There is no place in the Arrow–Debreu model for asymmetric information. The second welfare theorem, for example, relies on lump sum redistributions, i.e. redistributions that occur in advance of the market interactions. But if agents cannot be distinguished except through their market behaviour, then the redistribution must be a function of market behaviour. Rational agents, anticipating this, will distort their behaviour and the optimality of the redistribution will be lost.

Similarly, in the definition of equilibrium no agent takes into account what other agents know, for example about the state of nature. Thus it is quite possible in an Arrow–Debreu equilibrium for some ignorant agents to exchange valuable commodities for commodities indexed by states that other agents know will not occur. This problem received enormous attention in the finance literature, and some claim (see Grossman 1981) that it has been solved by extending the Arrow–Debreu definition of equilibrium to a ‘rational expectations equilibrium’ (Lucas 1972; see also Radner 1979). But this definition is itself suspect; in particular, it may not be implementable.

Even if rational expectations equilibrium (REE) were accepted as a visible notion of equilibrium, it could not come to grips with the most fundamental problems of asymmetric information. For like Arrow–Debreu equilibrium, in REE all trade is conducted anonymously through the market at given prices. Implicit in this definition is the assumption of large numbers of traders on both sides of every market. But what has come to be called the incentive problem in economics

revolves around individual or firm specific uncertainty, i.e. trade in commodities indexed by the names of the traders, which by definition involves few traders.

This brings us to another major riddle: how are agents supposed to get to equilibrium in the Arrow–Debreu model? The pioneers of general equilibrium never imagined that the economy was necessarily in equilibrium; Walras, for example, proposed an explicit tâtonnement procedure which he conjectured converged to equilibrium. But that idea is flawed in two respects: in general, it can be shown not to converge, and more importantly, it is an imaginary process in which no exchange is permitted until equilibrium is reached. This illustrates a grave shortcoming of any equilibrium theory, namely that it cannot begin to specify outcomes out of equilibrium. The major crisis of labour market clearing in the 1930s, and again recently, argues strongly that there are limits to the applicability of equilibrium analysis.

One is led naturally to consider market games, in which the outcomes are well- specified even when agents do not make their equilibrium moves. The most famous market game is Cournot's duopoly model, which has been extended to general equilibrium by Shapley–Shubik (1977). When there are a large number of agents of each type, the Nash equilibria of the Shapley–Shubik game give nearly identical allocations to the competitive allocations of Arrow–Debreu. This justifies (to first approximation) the price taking behaviour of the Arrow–Debreu agents. But note that the informational requirements of Nash equilibrium are at least twice that of Arrow–Debreu competitive equilibrium (each agent must know the aggregates of birds and offers on each market). It is also extremely interesting that trade takes place in the Shapley–Shubik game even if there is only one trader on each side of the market. Hence many problems in asymmetric information which have no place in the Arrow–Debreu model, because they involve too fine a specification of the commodities to be consistent with price taking, might be sensible in a market game context. Finally, it can be shown that REE is not consistent with the Shapley–Shubik game, or indeed with any continuous game.

We have indicated some of the ways in which it is possible to extend general equilibrium analysis to phenomena outside the scope of the Arrow–Debreu model, while at the same time preserving the neoclassical methodological premises of agent optimization, rational expectations, and equilibrium. It is important to note that these variations have extended the definition of equilibrium as well; this is most obvious in the case of market games, where Nash equilibrium replaces competitive equilibrium. All of the models have retained, on the other hand, more or less the same notion of rationality, sometimes at the cost of increasing the demands on the rationality of expectations. A great challenge for future general equilibrium models is how to formulate a sensible notion of bounded rationality, without destroying the possibility of drawing normative conclusions.

See Also

- ▶ Existence of General Equilibrium
- ▶ General Equilibrium
- ▶ Intertemporal Equilibrium and Efficiency
- ▶ Overlapping Generations Model of General Equilibrium

Bibliography

Arrow, K.J. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman, 507–532. Berkeley: University of California Press.

Arrow, K.J. 1953. *Le rôle des valeus boursières pour la répartition la meilleure des risques*, *Econométrie*, 41–48. Paris: Centre National de la Recherche Scientifique.

Arrow, K.J. 1969. The organization of economic activity: Issues pertinent to the choice of market vs nonmarket allocation. Reprinted in *Collected papers of Kenneth Arrow*, vol. II, 133–55. Cambridge, MA: Belknap Press.

Arrow, K.J., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.

Balasko, Y. 1986. *Foundations of the theory of general equilibrium*. New York: Academic Press.

Bewley, T. 1972. Existence of equilibria in economies with infinitely many commodities. *Journal of Economic Theory* 4: 514–540.
Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: L. Hachette.
Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19: 273–292.
Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*. New York: Wiley.
Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38: 387–392.
Debreu, G. 1974. Excess functions. *Journal of Mathematical Economics* 1: 15–21.
Debreu, G., and H. Scarf. 1963. A limit theorem on the core of an economy. *International Economic Review* 4: 235–246.
Dierker, E. 1974. *Topological methods in Walrasian economics*. Berlin: Springer.
Drèze, J. 1974. Investment under private ownership: Optimality, equilibrium, and stability. In *Allocation under uncertainty*, ed. J. Drèze. New York: Macmillan.
Duffie, D., and W. Shafer. 1985. Equilibrium in incomplete markets: I-A basic model of generic existence. *Journal of Mathematical Economics* 14 (3): 285–300.
Geanakoplos, J.D., and H.M. Polemarchakis. 1987. Existence, regularity, and constrained suboptimality of equilibrium with incomplete asset markets. In *Essays in honor of Kenneth J. Arrow*, ed. W. Heller, R. Starr, and D. Starrett, vol. 3. New York: Cambridge University Press.
Grossman, S. 1981. An introduction to the theory of rational expectations under asymmetric information. *Review of Economic Studies* 48: 541–560.
Grossman, S., and O. Hart. 1979. A theory of competitive equilibrium in stock market economies. *Econometrica* 47: 293–329.
Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11 (3): 418–433.
Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.
Koopmans, T.C., ed. 1951. *Activity analysis of production and allocation*. New York: Wiley.
Kuhn, H.W., and A.W. Tucker. 1951. Nonlinear programming. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman, 481–492. Berkeley: University of California Press.
Lange, O. 1942. The foundations of welfare economics. *Econometrica* 10: 215–228.
Lucas, R.E. 1972. Expectations and the neutrality of money. *Journal of Economic Theory* 4: 103–124.
McKenzie, L.W. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22: 147–161.
McKenzie, L.W. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27: 54–71.
Nash, J.F. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the USA* 36: 48–49.
Pareto, V. 1909. *Manuel d'économie politique*. Paris: Giard.
Radner, R. 1968. Competitive equilibrium under uncertainty. *Econometrica* 36 (1): 31–58.
Radner, R. 1979. Rational expectations equilibrium: Generic existence and the information revealed by prices. *Econometrica* 17: 655–678.
Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Samuelson, P.A. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66: 467–482.
Scarf, H. (with the collaboration of T. Hansen.). 1973. *The computation of economic equilibria*. New Haven: Yale University Press.
Shapley, L., and M. Shubik. 1977. Trade using one commodity as a means of payment. *Journal of Political Economy* 85: 937–968.
Smale, S. 1976. A convergent process of price adjustment and global Newton methods. *Journal of Mathematical Economics* 3: 107–120.
Sonnenschein, H. 1973. Do Walras' identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6: 345–354.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
von Neumann, J. 1928. Zur theorie der Gesellschaftsspiele. *Mathematische Annalen* 100: 295–320.
von Neumann, J. 1937. Uber ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. *Ergebnisse eines mathematischen Kolloquiums* 8: 73–83.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.
Walras, L. 1874–7. *Eléments d'économie politique pure*. Lausanne: L. Corbaz.

Art, Economics of

David Throsby

Abstract

The application of economic theory and analysis to problems in the performing arts (music, theatre, dance), the visual and literary arts and other art forms has expanded greatly over the last 30 years. A basic issue has been to identify

the ways in which artistic goods and services differ from other goods and services in the economy, thereby warranting particular attention. This article considers the economic analysis of demand and supply conditions in the arts, market structures (including factor markets) and a range of public policy issues in the area of cultural policy.

Keywords

Advertising; and cultural goods; Art, economics of; Baumol's disease; capital goods; Cultural goods as; charitable donations; And tax; Contingent valuation; Copyright; And 'fair use'; Cultural capital; Cultural goods and services; Defined; *Droit de suite*; Education; And arts demand; Herding; Heritage assets; Intellectual property; And cultural goods; Joint consumption; Market structure; Monopolistic competition; In live performing arts; Moral rights; Non-participant benefits; Non-pecuniary motives; In the arts; Not-for-profit firms; Output quality; Vs price as determinant of demand; Path dependence; Product differentiation; Superstars; Tax deductibility

JEL Classifications

Z11

The definition of art has been a philosophical conundrum for centuries, but there is probably a reasonable consensus on what comprises 'the arts'. These include the performing arts (music, dance, opera and theatre), the visual and plastic arts (painting, drawing, print-making, photography, sculpture, craft, and so on), the literary arts (poetry, fiction, drama, screenplays, and some forms of non-fiction such as biography), certain types of film, and some emerging practices such as video art that derive from new information and communications technologies. The application of economic theory and analysis across these various art forms comprises the discipline that has come to be known as cultural economics, although the ambit of this field has expanded in recent years to embrace wider economic questions relating to culture in an anthropological sense, such as the role of culture in economic development. Apart from some issues relating to the definition of cultural goods, this contribution does not deal with culture in the broader sense but rather is confined to the arts as defined above, and considers the conditions of demand, supply and exchange of artistic products, and some consequent issues for policy.

Characteristics of Cultural Goods

The goods and services produced by the arts, as well as some neighbouring commodities such as television programmes, video games and heritage services, can be called cultural goods and services. A fundamental question is whether such goods have unique characteristics that distinguish them as a commodity class from other goods and services in the economy. A reasonable definition of cultural goods attributes to them three necessary features: they require some input of human creativity in their manufacture; they possess or convey some symbolic meaning or messages; and they contain, at least potentially, some form of intellectual property. This definition extends to include a wide range of goods with only minor cultural content, such as fashion design, some forms of advertising, and some architectural services. Nevertheless, while there may be some blurring of boundaries at the cultural edges, there is little doubt that goods and services produced by the arts, as a subset of cultural goods, fit this definition nicely.

An alternative (or perhaps additional) definitional approach has been to portray cultural goods as embodying or giving rise to a form of value that lies beyond the reach of conventional economic assessment, and is not expressible (or is only imperfectly expressible) in market prices or in individual willingness-to-pay judgements. In the case of art works, such 'cultural' value might derive from ineffable aesthetic or spiritual qualities that such works of art are known to possess. These sources of value are only partially comprehensible within standard neoclassical price theory; indeed, they can be fully understood only by extending the analytical range to wider areas of economics, and beyond economics into other disciplines such as philosophy, psychology and aesthetics.

A further distinctive characteristic of the arts as consumption goods is that they are subject to the phenomenon of path dependence or, more specifically, rational addiction; that is, they are commodities for which an individual's present consumption depends on his or her past consumption, and patterns of demand tend to be cumulative. Although it is generally agreed that increased exposure to the arts in the past and the present will generate increased demand in the future (with consequent lessons for arts in education), this is hardly a sufficient condition for defining artistic goods, since a number of other commodities, not least addictive drugs, share a similar characteristic.

As economic commodities it is appropriate to categorize cultural goods as being capital goods, intermediate goods, or goods for final consumption. When classified as capital items (reusable goods whose services are combined with other inputs to produce further outputs), cultural goods have come to be known within economics as cultural capital, distinguished from other forms of capital by reference to either or both of the above definitions. This concept is especially relevant in the analysis of artworks and cultural heritage, where the interpretation of tangible or intangible cultural property as long-lasting assets created by the investment of resources, subject to depreciation unless properly maintained and yielding a rate of return over time, is readily understood.

It is important to note that cultural goods are generally very heterogeneous, suggesting that working in characteristics space may be a preferred way to analyse their demand and supply. For instance, demand for paintings can be thought of in Lancastrian terms as determined by the works' colour, size, style, school, and so on, and similar collections of characteristics can readily be imagined for other types of artistic commodities. Nevertheless, such heterogeneity does not vitiate the application of the tools of demand and supply analysis to the arts, as demonstrated further below.

Demand

A demand function for any type of artistic good or service could be expected to contain the usual sorts of explanatory variables: own price, price of substitutes, product quality characteristics and socio-demographic indicators relating to consumers' age, gender, income, education, and so forth. Within standard demand models, interest has focused on empirical questions: price and income elasticities, the relative importance of education and income, the cost of time, and the influence of quality aspects (to the extent that they can be measured). Results from a variety of art forms, time periods, geographical locations and data sources have varied widely, and even apparently plausible hypotheses, such as that the arts are a luxury good, have been by no means universally upheld. Nevertheless, the weight of evidence suggests, inter alia, that education is generally a more powerful predictor of arts demand than is income, and that output quality characteristics exert a strong influence on consumption patterns, perhaps overshadowing price as a determinant of demand behaviour in particular circumstances.

One topic of considerable interest in the demand for the performing arts is the emergence of so-called superstars, performers such as rock musicians and film actors whose incomes are greater than those of their competitors by a much larger differential than marginal productivity theory would suggest. Rosen (1981) attributed this phenomenon to two features of the demand for superstars' services. First, since consumers rationally prefer one good performance to two mediocre ones, particular types of services (such as rock music) are imperfect substitutes on the demand side, leading to convexity in sellers' returns and to a skewness in the distribution of earnings. Second, scale economies in joint consumption allow relatively few sellers to supply the entire market. Add to this the possible 'herding' behaviour of consumers, who follow the lead of others in making their demand decisions, and a plausible explanation as to why some performers command excessively high rents is obtained. Paradoxically, however, having broken away from the pack, superstars may finish up receiving *less* than their full earnings potential because some of their incremental contribution may have to be shared with employers, agents, managers and other beneficiaries of their superstardom.

Compared with the performing arts, the demand for art objects such as paintings – occurring in what is generally known as 'the art market' – raises some quite different questions. Durable works of art are sought by buyers not just for their aesthetic qualities but also because they are financial assets whose value may appreciate over time. Demand for paintings, prints, drawings, movable sculptures and other collectables such as silverware and rare books is readily separable into demand for art as a source of aesthetic gratification and demand for art as financial instrument. Both demands are affected by some of the same sorts of considerations – the reputation of the artist, the opinion of critics and market analysts, fashions in taste, past prices, and so on. At the same time other influences affect one or other aspect of demand specifically; for instance, demand for art as asset is constrained by some unattractive features of works of art as investments compared with alternative instruments, in particular their indivisibility, their illiquidity and their riskiness. In freely functioning markets, prices are expected to reflect all these influences, providing in equilibrium a means of balancing their respective importance. Since quite extensive and detailed data on prices in various art markets are available, a substantial econometric effort has been devoted to analysing price patterns across time and space for a wide range of types and styles of works of art. While much of this research yields results of interest only to art market specialists and connoisseurs – for example, do prices for paintings and prints by the same artist follow similar trends? – some of it addresses the more general issue of rates of return to art investment over time. Although contrary examples can be found, the general conclusion is that a collection of works of art will yield a lower return over the long term than a corresponding portfolio of stocks and bonds, the differential being attributable in part to the consumption services provided by the art for the period for which it is held.

Finally on the demand side, we can point to the demand for museum and heritage services. This demand includes attendances at art museums and heritage sites which provide private consumption experiences to the visitor, the specialist demand for conservation and restoration services provided by curators, art historians, and so on who staff the institutions concerned, and the demand for the public-good output of these cultural facilities, seen in the form of non-participant benefits accruing to the local and wider communities. With regard to direct visits to museums and sites, empirical experience suggests some price sensitivity, leading to arguments for free admission to publicly funded or operated facilities on the grounds that their educational and access benefits outweigh their potential for revenue raising. Nevertheless, in some instances, especially in the heritage field, revenue from visitors such as tourists is the only reliable source of ongoing funds for restoring or maintaining the facility concerned. However, regardless of the income-earning prospects of museum and heritage assets, the demand for their public-good output may well prove more decisive than the private-use demand for their services in rationalizing their existence in economic terms. In this respect demand estimation methods using stated preference techniques such as contingent valuation methods have proved useful in evaluating option, existence and bequest demands for these items of cultural capital and in quantifying willingness to pay for their services.

Supply

Artistic goods and services for final consumption are produced by a variety of types of enterprises ranging from single-person firms through small for-profit and not-for-profit companies to large corporate organizations in both private and public sectors. At the simplest end of this spectrum is the individual artist who produces goods or services for direct sale to the public – the visual artist selling paintings from her home, or the busker playing his saxophone in the shopping mall. From an economic viewpoint these artists can be seen as single-proprietor firms, probably unincorporated and subject to more than the usual vagaries of production, cost and market uncertainties that attend such producers elsewhere in the economy. Their labour time and their talent are likely to be their principal inputs, and their

production functions are likely to relate as much to the quality as to the quantity of their output. We return to the economic circumstances of individual artists below.

Across many fields in the arts – including opera, theatre, dance, classical music, jazz, independent film-making, small-scale literary publishing, contemporary visual art and craft, and so on – the predominant firm types, in terms of numbers of firms, are small and medium-sized enterprises, constituted on either a for-profit or a not-for-profit basis. Microeconomic theory offers straightforward means for characterizing the production and cost conditions under which all these firms operate, with differences according to specific features of the various industries. For example, in the performing arts the unit of output in both production and cost function estimations is generally taken as paid attendances, in a manner similar to the way output is measured in other service-providing firms such as hospitals and universities. Standard functional forms can be used to investigate elasticities of output with respect to various inputs, economies of scale and scope, technical and allocative efficiencies, and productivity growth.

While production and cost conditions may be expected to be similar for these firms whether they are profit-oriented or otherwise, the structure and behaviour of for-profit and not-for-profit firms will differ markedly. Much attention in the economics of the arts has been focused on the latter because of the prevalence of not-for-profit firms at the 'serious' end of the artistic spectrum, producing innovative output or work which, though judged artistically worthy, does not appeal to a mass audience. Not only is there insufficient demand to sustain commercial production of this sort of work, but also the motives of the firms producing it are artistic rather than pecuniary. They can therefore be modelled as constrained maximizers of output quality (and possibly of the quantity of output as well if they wish to spread their art to as wide an audience as possible); the constraint is a break-even restriction whereby earned plus unearned revenue must at least cover costs over some specified period. Other model specifications have also been investigated, for example incorporating an objective of maximizing revenues from sponsorship and donations.

An issue of continuing interest in the economics of the performing arts is that of productivity lag, first identified by Baumol and Bowen (1966) and subsequently labelled 'Baumol's disease' or 'the cost disease'. Essentially the hypothesis states that labour productivity in the live arts remains static over time – it still takes the same number of workers the same amount of time to perform *Hamlet* today as it did in Shakespeare's day. In a two-sector model in which one sector suffers from this technological disadvantage, wage rises in the productive sector are transmitted to the stagnant sector, causing a widening gap in the latter between revenues and costs, since firms in the stagnant sector cannot cover wage rises with improved labour productivity. Applying this to the live arts, Baumol and Bowen predicted that performing firms would have to access increasing levels of non-box-office revenue over time in order to stay in business. Empirical studies of this phenomenon have confirmed that costs of live performances have indeed risen as the model implies, but that the impact of these cost increases on firms has been somewhat muted; most performing companies have been able to mitigate the effects of slow productivity growth through a variety of strategies, including tapping new sources of unearned revenue, exploiting the potential of new recording and distribution technologies, expanded ancillary activities such as merchandising, and so on.

Finally in this section we turn to large-scale production in the arts. There are certainly some not-for-profit firms in the arts with multi-million dollar budgets, including major art museums, the world's principal opera companies and symphony orchestras, national theatre companies in several countries, and so on. In almost all cases some level of public funding is involved, together with significant levels of private-sector support from foundations, corporations and individual donors to supplement box-office revenue. In some countries these large-scale enterprises are government business undertakings, subject to varying degrees of independence or control in their governance

and their operational decision-making. However, the majority of large-scale producers of artistic goods are profit-seeking firms operating in commercial markets where complex production processes are required and/or where substantial scale economies exist. These firms include theatre companies staging popular shows, commercial and independent film producers, music publishers, record companies, major book publishers, art auction houses and so on. Taken together, these firms form a significant component (measured in terms of value of output) of the so-called creative or copyright industries, terms reflecting two of the necessary characteristics of cultural goods discussed earlier. From an economic point of view, these industries are notable for their peculiar contractual arrangements that reflect, among other things, the inherent uncertainties that attend every stage of artistic production processes whereby 'nobody knows' what the quality or market potential of the final product will be (Caves 2000).

Market Structures

It is perhaps surprising that there is little in the industrial organization literature dealing with structure, conduct and performance in the arts. There are many interesting questions concerning competition, market efficiency and pricing behaviour in the arts that await the attention of economists. As may be evidenced from the preceding section, the range of market structures in the arts is quite wide, providing considerable scope for empirical investigation.

At one extreme can be found instances of almost atomistic competition, as in the so-called primary market for visual art. Here there are many small producers, mostly individual artists selling on their own or through small local galleries, art fairs, and so on. Although the product is not exactly homogeneous, buyers tend to be not very discriminating, and prices may well be competed down to little more than cost of production plus some modest return to labour. Moving further across the market structure spectrum, we can suggest that the live performing arts in medium-to-large towns and cities show some evidence of monopolistic competition: a relatively large number of small firms competing through product differentiation and other non-price strategies for customers drawn from a single pool. Higher levels of concentration appear in other areas of the arts, especially in local markets for live performance characterized by one or two dominant firms when close substitutes are not available; the markets for opera or orchestral music in a given city may be examples. In all of the above cases, market conditions affect the pricing and output decisions of participating firms. Given that non-pecuniary motives play an important role in influencing the behaviour of economic agents in the arts, the competitive outcomes in the markets discussed might be expected to diverge somewhat from those predicted under more conventional conditions.

Factor Markets

The input into artistic production processes that provides the unique qualities of artistic goods and services is, of course, the creative labour of artists themselves. Labour markets in the arts have been widely studied in both theoretical and empirical terms in an effort to understand whether and in what ways they differ from conventional labour markets. A principal finding relates again to the non-pecuniary motives for artistic production. Artists in general do not regard work as a chore whose only purpose is to earn an income. Rather, their commitment to making art means that they have a positive preference for working at their chosen profession, and empirical evidence indicates that they often forgo lucrative alternative employment in order to spend more time pursuing their creative work. This can be modelled as a time allocation problem where the worker has to choose between preferred but less remunerative work in the arts on the one hand and better-paid but less desired non-arts work on the other. The choice is subject to a minimum-income constraint, necessary to prevent starvation, a condition often romantically associated with artists but rarely observed in practice. Such a 'work preference' model of labour supply yields predictions of behaviour at

variance with the usual textbook construct – for example, a wage rise in the non-arts occupation may induce *less* work in that occupation because it enables more time to be devoted to the arts, a phenomenon akin to the backward-bending supply curve of labour in the conventional model.

The generally low levels of average earnings available from artistic practice mean that arts labour markets are characterized by ubiquitous multiple job-holding and much fluidity in career paths. The distribution of earnings across any population of arts workers is almost always skewed towards the lower end. Some attention has been paid to the role of risk in affecting entry and exit decisions in arts labour markets. Given the superstar phenomenon noted above, where extremely high incomes are earned by very few, some writers have portrayed these labour markets as winner-take-all lotteries to which artists submit themselves willingly. An alternative explanation of persistent labour market participation when expected monetary returns are low lies in the supposition that artists earn a sufficient level of psychic income to offset the meagre levels of their pecuniary rewards.

Turning to capital markets, we note simply that a similar psychic component may be present in rewarding suppliers of capital to the arts. For example, investors willing to back a theatre company putting on a new show may perhaps do so in expectation that the show will be a hit and they will earn a handsome return on their investment; however, a more plausible explanation for such a risky decision may be that these donors are motivated by a love of the theatre and hence that their satisfaction will derive largely if not entirely from the psychic rewards from helping to make it happen. Indeed, much private capital flows to the arts not as investments or loans but as untied donations with no strings attached, as discussed further below.

Policy Issues

Government provision of financial assistance to the arts is widespread across the developed world, though the extent of intervention varies substantially between countries and between jurisdictions within countries. It is not clear whether such assistance is in accord with the wishes of voters or whether it is a case of imposed preferences whereby the arts are seen by governments as a merit good. It is also entirely possible that public subsidies to the arts are consistent with the restoration of Pareto optimality in an economy subject to market failure, if it is indeed the case that the arts give rise to public goods or positive externalities. Some economists remain sceptical of the latter proposition on empirical rather than theoretical grounds, and there is as yet not a great deal of evidence to resolve the issue one way or the other. In these circumstances more attention has been focused on the appropriate means for intervention once a normative rationale is accepted. The instruments governments have at their disposal include public-sector provision of artistic services (for example, through public art galleries); direct subsidies to cultural production or consumption; indirect support through the tax system; regulation; provision of information; assistance through the education system; and so on. An issue of considerable interest is the specification of optimal decision rules for allocation of public financing among competing avenues of artistic activity, a process apparently driven as much by rent-seeking or political expediency as by the pursuit of economic efficiency.

The use of the tax system as a means of providing assistance has been of particular significance to the arts, especially via the tax deductibility allowed to philanthropic donors who give money to not-for-profit performing companies, museums, galleries, and so on. Such giving is likely to be motivated by a desire to secure the sorts of public-good benefits of the arts mentioned earlier, in circumstances where direct government support is regarded as inadequate. In some countries, most notably the United States, the cost of indirect support for the arts, measured in terms of tax revenue forgone, greatly exceeds the amount of direct financing by the public sector. Given that governments can manipulate the incentives facing donors by changing marginal tax rates, by raising or lowering thresholds and ceilings on allowable donations, and so on, much interest has focused on elasticities of giving with respect to variables such as the tax

price. The critical issue from a policy viewpoint is whether the price elasticity is greater or less than unity in absolute terms, since a price elastic response would imply that lowering the tax price would increase recipients' revenue by more than the tax receipts forgone. However, despite many empirical studies, no clear consensus as to the size of these elasticities has emerged. Other policy issues of concern in this field include whether increased government support for the arts crowds out or crowds in private donations, and whether it is good or bad policy to use an instrument that allows private individuals to direct the allocation of public resources via their charitable-giving decisions.

One way in which public policy can assist the functioning of markets in the arts is via the creation and enforcement of property rights in artistic goods and services. Efficient copyright regimes aim to facilitate public access to information, at the same time as allowing creators to regulate the use of their work and to capture remuneration that would otherwise be lost to piracy, free-riding, unauthorized commercial exploitation, and the like. While often seen as a purely legal matter, copyright has a number of economic implications for the arts. In particular, artistic output in the form of literary works, paintings, photographs, musical compositions, and so forth can generally be reproduced at low or negligible cost, and in the absence of copyright protection their price would be driven down to marginal cost, so reducing or eliminating the incentive to the artist to create further output. Nevertheless, some exceptions to universal copyright coverage exist, for example in the 'fair use' provisions of copyright law, which allow free access for certain scholarly or public-interest purposes, or where high transactions costs of enforcement outweigh the potential gains to the rights holder. Other intellectual property issues of interest to economists include the market effects of moral rights (the rights that artists have over attribution and integrity of their works) and, in the visual arts, the phenomenon of *droit de suite* (the payment of a royalty to the artist or his or her heirs each time a given work is resold).

An area of growing importance in policy terms in recent years has been the role of the arts in urban and regional development. This role may be evident in a specific sense, for example in the impact of an arts festival on the local economic base, or in the use of community arts projects to engage and motivate disaffected youth in areas of high unemployment. In a wider context, the creative industries may be seen as a source of new enterprise, income growth and employment creation in depressed industrial regions. Empirical studies have looked at the impact of arts events, facilities, and so forth on a local or regional economy, and at the more general contribution that the arts industries make to economic activity, as a basis for policy formulation in a field increasingly engaging the attention of governments at both national and local levels.

Public policy towards the arts, heritage, the creative industries, cultural trade, and so forth can be gathered together under the somewhat fuzzy heading of 'cultural policy'. Given the significant economic content of all of these areas, it can be expected that economic theory and analysis will continue to make an important contribution to policy-making in this field in the future.

Further Reading

Recent surveys of the economics of the arts include Throsby (1994), Blaug (2001) and Ginsburgh (2001). Major contributions to the literature on the economics of the arts from the mid-1960s to the mid-1990s are collected together in Towse (1997). A broader view of cultural economics is contained in Throsby (2001). An accessible account of the principal topics in contemporary cultural economics is provided in Towse (2003), while a comprehensive research-oriented coverage of the economics of art and culture is contained in Ginsburgh and Throsby (2006).

See Also

- ▶ Books, Economics of
- ▶ Intellectual Property
- ▶ Music Markets, Economics of
- ▶ Superstars, Economics of

Bibliography

Baumol, W., and W.G. Bowen. 1966. *Performing arts: The economic dilemma*. New York: Twentieth Century Fund.
Blaug, M. 2001. Where are we now in cultural economics? *Journal of Economic Surveys* 15: 123–143.
Caves, R. 2000. *Creative industries: contracts between art and commerce*. Cambridge, MA: Harvard University Press.
Ginsburgh, V. 2001. Economics of arts and culture. In *International encyclopaedia of the social and behavioural sciences*, ed. N. Smelser and P. Baltes. Amsterdam: Elsevier.
Ginsburgh, V., and D. Throsby, eds. 2006. *Handbook of the economics of art and culture*. Amsterdam: Elsevier/North Holland.
Rosen, S. 1981. The economics of superstars. *American Economic Review* 71: 845–858.
Throsby, D. 1994. The production and consumption of the arts: a view of cultural economics. *Journal of Economic Literature* 32: 1–29.
Throsby, D. 2001. *Economics and culture*. Cambridge: Cambridge University Press.
Towse, R., eds. 1997. *Cultural economics: The arts, the heritage and the media industries*. Vol. 2. Cheltenham: Edward Elgar.
Towse, R., eds. 2003. *A handbook of cultural economics*. Cheltenham: Edward Elgar.

Artificial Neural Networks

Chung-Ming Kuan

Abstract

Artificial neural networks (ANNs) constitute a class of flexible nonlinear models designed to mimic biological neural systems. In this article we introduce ANN using familiar econometric terminology and provide an overview of the ANN modelling approach and its implementation methods.

Keywords

ARMA models; Artificial neural networks; Back propagation algorithm; BAYESIAN information criterion; Distributed lags; Feedforward neural networks; Learning; Model misspecification; Newton algorithm; Nonlinear least squares (NLS) method; Nonlinear models; Nonparametric functional analysis; Nonparametric methods; Predictive stochastic complexity; Recurrent neural networks; Stochastic approximation method

JEL Classification

C45

Introduction

Artificial neural networks (ANNs) constitute a class of flexible nonlinear models designed to mimic biological neural systems. Typically, a biological neural system consists of several layers, each with a large number of neural units (neurons) that can process the information in a parallel manner. The models with these features are known as ANN models. Such models can be traced back to the simple input–output model of McCulloch and Pitts (1943) and the 'perceptron' of Rosenblatt (1958). The early yet simple ANN models, however, did not receive much attention because of their limited applicability and also because of the limitation of computing capacity at that time. In seminal works, Rumelhart, McClelland and PDP Research Group (1986b) and McClelland, Rumelhart and PDP Research Group (1986) presented the new developments of ANN, including more complex and flexible ANN structures and a new network learning method. Since then, ANN has become a rapidly growing research area.

As far as model specification is concerned, ANN has a multi-layer structure such that the middle layer is built upon many simple nonlinear functions that play the role of neurons in a biological system. By allowing the number of these simple functions to increase indefinitely, a multi-layered ANN is capable of approximating a large class of functions to any desired degree of accuracy, as shown in, for example, Cybenko (1989), Funahashi (1989), Hornik, Stinchcombe and White (1989, 1990), and Hornik (1991, 1993). From an econometric perspective, ANN

can be applied to approximate the unknown conditional mean (median, quantile) function of the variable of interest without suffering from the problem of model misspecification, unlike parametric models commonly used in empirical studies. Although nonparametric methods, such as series and polynomial approximators, also possess this property, they usually require a larger number of components to achieve similar approximation accuracy (Barron 1993). ANNs are thus a parsimonious approach to nonparametric functional analysis.

ANNs have been widely applied to solve many difficult problems in different areas, including pattern recognition, signal processing, and language learning. Since White (1988), there have also been numerous applications of ANN in economics and finance. Unfortunately, the ANN literature is not easy to penetrate, so it is hard for applied economists to understand why ANN works and how it can be implemented properly. Fortunately, while the ANN jargon originated from cognitive science and computer science, they often have econometric interpretations. For example, a 'target' is, in fact, a dependent variable of interest, an 'input' is an explanatory variable, and network 'learning' amounts to the estimation of unknown parameters in a network. The purpose of this article is thus twofold. First, it introduces ANN using familiar econometric terminology and hence serves to bridge the gap between the fields of ANN and economics. Second, it provides an overview of ANN modelling approach and its implementation methods. For an early review of ANN from an econometric perspective, we refer to Kuan and White (1994).

This article proceeds as follows. We introduce various ANN model specifications and the choices of network functions in section "ANN model specifications". We present the 'universal approximation' property of ANN in section "ANN as an universal approximator". Model estimation and model complexity regularization are discussed in section "Implementation of ANNs". Section "Concluding remarks" concludes.

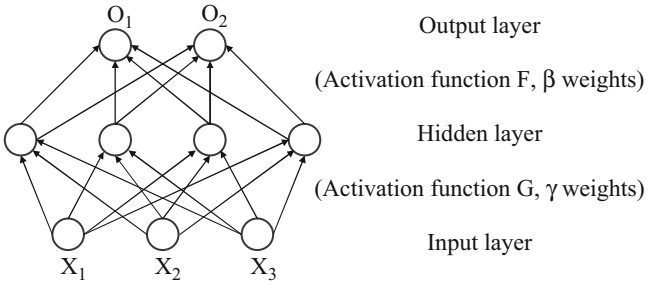
ANN Model Specifications

Let Y denote the collection of n variables of interest with the t-th observation $\mathbf{y}_t$ ($n \times 1$) and X the collection of m explanatory variables with the t-th observation $\mathbf{x}_t$ ($m \times 1$). In the ANN literature, the variables in Y are known as *targets* or *target variables*, and the variables in X are *inputs* or *input variables*. There are various ways to build an ANN model that can be used to characterize the behavior of $\mathbf{y}_t$ using the information contained in the input variables $\mathbf{x}_t$. In this section, we introduce some network architectures and the functions that are commonly used to build an ANN.

Feedforward Neural Networks

We first consider a network with an *input* layer, an *output* layer, and a *hidden* layer in between. The input (output) layer contains m input units (n output units) such that each unit corresponds to a particular input (output) variable. In the hidden layer, there are q hidden units connected to all input and output units; the strengths of such connections are labelled by (unknown) parameters known as the network *connection weights*. In particular, $\gamma_h = (\gamma_{h.1},\ldots,\gamma_{h,m})'$ denotes the vector of the connection weights between the h-th hidden unit and all m input units, and $\boldsymbol{\beta}_j = (\beta_{j,1};\ldots;\beta_{j,q})'$ denotes the vector of the connection weights between the j-th output unit and all q hidden units. An ANN in which the sample information (signals) are passed forward from the input layer to the output layer without feedback is known as a *feedforward neural network*. Figure 1 illustrates the architecture of a three-layer feedforward network with three input units, four hidden units and two output units.

This multi-layered structure of a feedforward network is designed to function as a biological neural system. The input units are the neurons that receive the information (stimuli) from the outside environment and pass them to the neurons in a middle layer (that is, hidden units). These neurons then transform the input signals to generate neural signals and forward them to the neurons in the output layer. The output neurons in turn generate signals that determine the action to be



Artificial Neural Networks,
Fig. 1 A feedforward network with three input units, four hidden units and two output units

taken. Note that all information from the units in one layer is processed simultaneously, rather than sequentially, by the units in an 'upper' layer. (This concept, also known as *parallel processing* or *massive parallelism*, differs from the traditional concept of sequential processing and has led to a major advance in designing computer architecture.)

Formally, the input units receive the information x_t and send to all hidden units, weighted by the connection weights between the input and hidden units. This information is then transformed by the *activation function* G in each hidden unit. That is, the h-th hidden unit receives $\mathbf{x}_t'\gamma_h$ and transforms it to $G(\mathbf{x}_t'\gamma_h)$. The information generated by all hidden units is further passed to the output units, again weighted by the connection weights, and transformed by the activation function F in each output unit. Hence, the j-th output unit receives $\sum_{h=1}^{q}\beta_{j,h}G(\mathbf{x}_t'\gamma_h)$ and transforms it into the network output:

$$o_{t,j} = F\left(\sum_{h=1}^{q}\beta_{j,h}G(\mathbf{x}_t'\gamma_h)\right),\ j = 1, \ldots, n. \quad (1)$$

The output O_j is used to describe or predict the behaviour of the j-th target Y_j.

In practice, it is typical to include a constant term, also known as the *bias* term, in each activation function in (1). That is,

$$o_{t,j} = F\left(\beta_{j,0} + \sum_{h=1}^{q}\beta_{j,h}G(\gamma_{j,0} + \mathbf{x}_t'\gamma_h)\right), j = 1, \ldots, n, \quad (2)$$

where $\gamma_{h,0}$ is the bias term in the h-th hidden unit and $\beta_{j,0}$ is the bias term in the j-th output unit. A constant term in each activation function adds flexibility to hidden-unit and output-unit responses (activations), in a way similar to the constant term in (non)linear regression models. Note that when there is no transformation in the output units, F is an identity function (that is, $F(a) = a$) so that

$$o_{t,j} = \beta_{j,0} + \sum_{h=1}^{q}\beta_{j,h}G(\gamma_{j,0} + \mathbf{x}_t'\gamma_h),\ j = 1, \ldots, n. \quad (3)$$

It is also straightforward to construct networks with two or more hidden layers. For simplicity, we will focus on the three-layer networks with only one hidden layer.

While parametric econometric models are typically formulated using a given function of the input x_t, the network (2) is a class of flexible nonlinear functions of x_t. The exact form of a network model depends on the activation functions (F and G) and the number of hidden units (q). In particular, the network function in (3) is an affine transformation of G and hence may be interpreted as an expansion with the 'basis' function G.

The networks (2) and (3) can be further extended. For example, one may construct a network in which the input units are connected not only to the hidden units but also directly to the output units. This leads to networks with *short-cut* connections. Corresponding to (2), the outputs of a feedforward network with short cuts are

$$o_{t,j} = F\left(\beta_{j,0} + \mathbf{x}_t'\boldsymbol{\alpha}_j \sum_{h=1}^{q} \beta_{j,h} G\left(\mathbf{x}_t'\gamma_{\mathrm{h}}\right)\right), \quad j = 1, \ldots, n,$$

where α_j is the vector of connection weights between the output and input units, and, corresponding to (3), the outputs are

$$o_{t,j} = \beta_{j,0} + \mathbf{x}_t'\boldsymbol{\alpha}_j \sum_{h=1}^{q} \beta_{j,h} G\left(\gamma_{j,0} + \mathbf{x}_t'\gamma_{\mathrm{h}}\right), \quad j = 1, \ldots, n.$$

Figure 2 illustrates the architecture of a feedforward network with two input units, three hidden units, one output unit and short-cut connections. Thus, parametric econometric models may be interpreted as feedforward networks with short-cut connections but no hidden-layer connections. The linear combination of hidden-unit activations, $\sum_{h=1}^{q} \beta_{j,h} G\left(\gamma_{h,0} + \mathbf{x}_t'\gamma_{\mathrm{h}}\right)$, in effect characterizes the nonlinearity not captured by the linear function of $\mathbf{x}_t$.

Recurrent Neural Networks

From the preceding section we can see that there is no 'memory' device in feedforward networks that can store the signals generated earlier. Hence, feedforward networks treat all sample information as 'new'; the signals in the past do not help to identify data features, even when sample information exhibits temporal dependence. As such, a feedforward network must be expanded to a large extent so as to represent complex dynamic patterns. This causes practical difficulty because a large network may not be easily implemented. To utilize the information from the past, it is natural to include lagged target information $\mathbf{y}_{t-k}$, $k = 1,\ldots, s$, as input variables, similar to linear AR and ARX models in econometric studies. Yet such networks do not have any built-in structure that can 'memorize' previous neural responses (transformed sample information). The so-called *recurrent neural networks* overcome this difficulty by allowing internal feedbacks and hence are especially appropriate for dynamic problems.

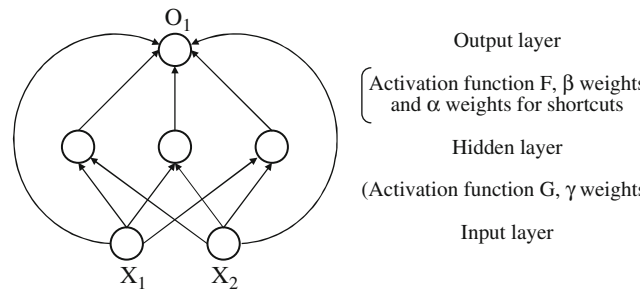
Jordan (1986) first introduced a recurrent network with feedbacks from output units. That is, the output units are connected to input units but with *time delay*, so that the network outputs at time $t-1$ are also the input information at time t. Specifically, the outputs of a Jordan network are

$$o_{t,j} = F\left(\beta_{j,0} + \sum_{h=1}^{q} \beta_{j,h} G\left(\gamma_{j,0} + \mathbf{x}_t'\gamma_{\mathrm{h}} + \mathbf{o}_{t-1}'\boldsymbol{\delta}_h\right)\right), \quad j = 1, \ldots, n, \tag{4}$$

where δ_h is the vector of the connection weights between the h-th hidden unit and the input units that receive lagged outputs $\mathbf{o}_{t-1} = (o_{t-1,1},\ldots, o_{t-1,n})'$. The network (4) can be further extended to allow for more lagged outputs $\mathbf{o}_{t-2}$, $\mathbf{o}_{t-3}$,.. ..

Similarly, Elman (1990) considered a recurrent network in which the hidden units are connected to input units with time delay. The outputs of an Elman network are:

$$\begin{aligned} o_{t,j} &= F\left(\beta_{j,0} + \sum_{h=1}^{q} \beta_{j,h} a_{t,h}\right), \quad j = 1, \ldots, n, \\ a_{t,h} &= G\left(\gamma_{h,0} + \mathbf{x}_t'\gamma_{\mathrm{h}} + \mathbf{a}_{t-1}'\boldsymbol{\delta}_h\right), h = 1, \ldots, q, \end{aligned} \tag{5}$$



Artificial Neural Networks, Fig. 2 A feedforward neural network with short cuts

where $\mathbf{a}_{t-1} = (a_{t-1,1},\ldots, a_{t-1,q})'$ is the vector of lagged hidden-unit activations, and $\boldsymbol{\delta}_h$ here is the vector of the connection weights between the h-th hidden unit and the input units that receive lagged hidden-unit activations $\mathbf{a}_{t-1}$. The network (5) can also be extended to allow for more lagged hidden-unit activations $\mathbf{a}_{t-2}$, $\mathbf{a}_{t-3}$, Figure 3 illustrates the architectures of a Jordan network and an Elman network.

From (4) and (5) we can see that, by recursive substitution, the outputs of these recurrent networks can be expressed in terms of current and all past inputs. Such expressions are analogous to the distributed lag model or the AR representation of an ARMA model (when the inputs are lagged targets). Thus, recurrent networks incorporate the information in the past input variables without including all of them in the model. By contrast, a feedforward network requires a large number of inputs to carry such information. Note that the Jordan network and the Elman network summarize past input information in different ways and hence have their own merits. When the previous 'location' of a network is crucial in determining the next move, as in the design of a robot, a Jordan network seems more appropriate. When the past internal neural responses are more important, as in language learning problems, an Elman network may be preferred.

Choices of Activation Function

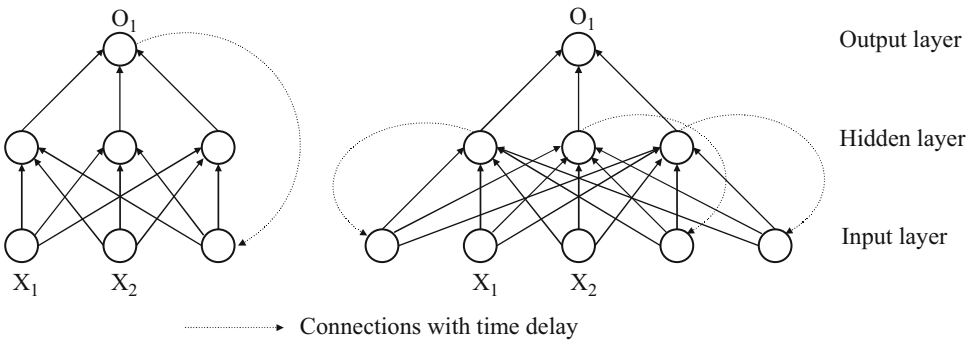
As far as model specifications are concerned, the building blocks of an ANN model are the activation functions F and G. Different choices of the activation functions result in different network models. We now introduce some activation functions commonly employed in empirical studies.

Recall that the hidden units play the role of neurons in a biological system. Thus, the activation function in each hidden unit determines whether a neuron should be turned on or off. Such an on/off response can be easily represented using an indicator (threshold) function, also known as a *heaviside* function in the ANN literature, that is,

$$G\left(\gamma_{h,0} + \mathbf{x}_t'\gamma_{\mathrm{h}}\right) = \begin{cases} 1, & \text{If } \gamma_{h,0} + \mathbf{x}_t'\gamma_{\mathrm{h}} \geq c, \\ 0, & \text{If } \gamma_{h,0} + \mathbf{x}_t'\gamma_{\mathrm{h}} < c, \end{cases}$$

where c is a pre-determined threshold value. That is, depending on the strength of connection weights and input signals, the activation function G will determine whether a particular neuron is on $\left(G\left(\gamma_{h,0} + \mathbf{x}_t'\gamma_{\mathrm{h}}\right) = 1\right)$ or off $\left(G\left(\gamma_{h,0} + \mathbf{x}_t'\gamma_{\mathrm{h}}\right) = 1\right)$.

In a complex neural system, neurons need not have only an on/off response but may be in an intermediate position. This amounts to allowing the activation function to assume any value between 0 and 1. In the ANN literature, it is common to choose a *sigmoid* (*S*-shaped) and *squashing* (bounded) function. In particular, if the input signals are 'squashed' between 0 and 1, the activation function is understood as a smooth counterpart of the indicator function. A leading example is the logistic function:



Artificial Neural Networks, Fig. 3 Recurrent neural networks: Jordan (*left*) and Elman (*right*)

$$G(\gamma_{h,0} + \mathbf{x}_t'\gamma_h) = \frac{1}{1 + \exp(-[\gamma_{h,0} + \mathbf{x}_t'\gamma_h])},$$

which approaches 1 (zero) when its argument goes to infinity (negative infinity). Hence, the logistic activation function generates a partially on/off signal based on the received input signals.

Alternatively, the hyperbolic tangent (tanh) function, which is also a sigmoid and squashing function, can serve as an activation function:

$$G(\gamma_{h,0} + \mathbf{x}_t'\gamma_h) = \frac{\exp(\gamma_{h,0} + \mathbf{x}_t'\gamma_h) - \exp(-[\gamma_{h,0} + \mathbf{x}_t'\gamma_h])}{\exp(\gamma_{h,0} + \mathbf{x}_t'\gamma_h) + \exp(-[\gamma_{h,0} + \mathbf{x}_t'\gamma_h])}.$$

Compared with the logistic function, this function may assume negative values and is bounded between -1 and 1. It approaches 1 (-1) when its argument goes to infinity (minus infinity). This function is more flexible because the negative values, in effect, represent 'suppressing' signals from the hidden unit. See Fig. 4 for an illustration of the logistic and tanh functions. Note that for the logistic function G, a re-scaled function $\tilde{G}$ such that $\tilde{G}(a) = 2G(a) - 1$ also generates values between -1 and 1 and may be used in place of the tanh function. (A choice of the activation function in classification problems is the so-called *radial basis function*. We do not discuss this choice because its argument is not an affine transformation of inputs and hence does not fit in our framework here. Moreover, the networks with this activation function provide only *local* approximation to unknown functions, in contrast with the approximation property discussed in section "ANN as an universal approximator".)

The aforementioned activation functions are chosen for convenience because they are differentiable everywhere and their derivatives are easy to compute. In particular, when G is the logistic function,

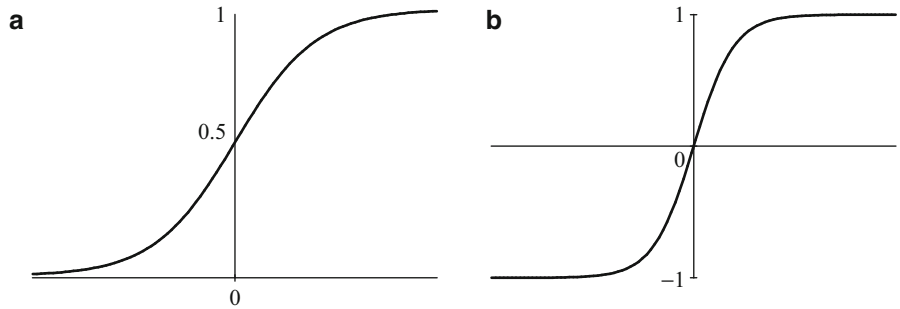
$$\frac{dG(a)}{da} = G(a)[1 - G(a)];$$

when G is the tanh function,

$$\frac{dG(a)}{da} = \left[\frac{2}{\exp(a) + \exp(-a)}\right]^2 = \mathrm{sech}^2(a).$$

These properties facilitate parameter estimation, as will be seen in section "Model estimation". Nevertheless, these functions are not necessary for building proper ANNs. For example, smooth cumulative distribution functions, which are sigmoidal and squashing, are also legitimate candidates for activation function. In section "ANN as an universal approximator", it is shown that, as far as network approximation property is concerned, the activation function in hidden units does not even have to be sigmoidal, yet boundedness is usually required. Thus, sine and cosine functions can also serve as an activation function.

As for the activation function F in the output units, it is common to set it as the identity function so that the outputs of (3) enjoy the freedom of assuming any real value. This choice suffices for the network approximation property discussed in section "ANN as an universal approximator".



Artificial Neural Networks, Fig. 4 Activation functions: logistic (*left*) and tanh (*right*)

When the target is a binary variable taking the values zero and one, as in a classification problem, F may be chosen as the logistic function so that the outputs of (2) must fall between 0 and 1, analogous to a logit model in econometrics.

ANN as an Universal Approximator

What makes ANN a useful econometric tool is its *universal approximation* property, which basically means that a multi-layered ANN with a large number of hidden units can well approximate a large class of functions. This approximation property is analogous to that of nonparametric approximators, such as polynomials and Fourier series, yet it is not shared by parametric econometric models.

To present the approximation property, we consider the network function element by element. Let $f_{G,q} : \mathbb{R}^m \times \Theta_{m,q} \to \mathbb{R}$ denote the network function with q hidden units, the output activation function F being the identity function, and the hidden-unit activation function G, that is,

$$f_{G,q}(\mathbf{x};\boldsymbol{\theta}) = \beta_0 + \sum_{h=1}^{q} \beta_h G\left(\gamma_{h,0} + \mathbf{x}'\gamma_h\right),$$

as in (3), where $\Theta_{m,q}$ is the parameter space whose dimension depends on m and q, and $\boldsymbol{\theta} \in \Theta_{m,q}$ (note that the subscripts m and q for $\boldsymbol{\theta}$ are suppressed). Given the activation function G, the collection of all $f_{G,q}$ functions with different q is:

$$\begin{aligned}\mathcal{F}_G = \bigcup_{q=1}^{\infty} \Big\{ f_{G,q} : f_{G,q}(\mathbf{x};\boldsymbol{\theta}) \\ = \beta_0 + \sum_{h=1}^{q} \beta_h \mathrm{G}\left(\gamma_{h,0} + \mathrm{x}'\gamma_h\right) \Big\}\end{aligned}$$

when the union is taken up to a finite number N, the resulting collection is denoted as $\mathcal{F}_G^N$. Intuitively, $\mathcal{F}_G$ is capable of functional approximation because $f_{G,q}$ can be viewed as an expansion with the 'basis' function G and hence is similar to a nonparametric approximator.

More formally, we follow Hornik (1991) and consider two measures of the closeness between functions. First define the *uniform distance* between functions f and g on the set K as

$$d_K(f,g) = \sup_{\mathbf{x}\in K} |f(\mathbf{x}) - g(\mathbf{x})|.$$

Let K denote a compact subset in $\mathbb{R}^m$ and $C(K)$ denote the space of all continuous functions on K. Then, when the activation function G is continuous, bounded and nonconstant, the collection $\mathcal{F}_G$ is dense in $C(K)$ for all K in $\mathbb{R}^m$ in terms of d_K (Theorem 2 of Hornik 1991). (Hornik 1991, considered the network without the bias term in the output unit, that is, $\beta_0 = 0$. Yet as long as G is not a constant function, all the results in Hornik 1991, carry over; see Stinchcombe and White 1998, for details.) That is, for any function g in $C(K)$ and any $\varepsilon > 0$, there is a network function $f_{G,q}$ in $\mathcal{F}_G$ such that $d_K(f_{G;q} - g) < \varepsilon$. As $\mathcal{F}_G^N$ is not dense in $C(K)$ for any finite number N, this result shows that any continuous function can be approximated arbitrarily well on compacta by a three-layered feedforward network $f_{G,q}$, provided that q, the number of hidden units, is sufficiently large.

Taking x as random variables, defined in the probability space with the probability measure $\mathbb{P}$, we consider the L_r-*norm* of $f(\mathbf{x})-g(\mathbf{x})$:

$$\|f-g\|_r \left(\int_{\mathbb{R}^m} |f(\mathbf{x}) - g(\mathbf{x})|^r d\mathbb{P}(\mathbf{x}) \right)^{1/r},$$

$1 \le r < \infty$. For $r = 2$ ($r = 1$), this is the well-known measure of mean squared error (mean absolute error). Then, when the activation function G is bounded and nonconstant, the collection $\mathcal{F}_G$ is dense in the $\mathrm{L_r}$ space (Theorem 1 of Hornik 1991). That is, any function g (with finite $\mathrm{L_r}$-norm) can also be well approximated by a three-layered feedforward network $f_{G,q}$ in terms of $\mathrm{L_r}$-norm when q is sufficiently large.

It should be emphasized that the universal approximation property of a feedforward network hinges on the three-layered architecture and the number of hidden units, but not on the activation function per se. As stated above, the activation function in the hidden unit can be a general bounded function and does not have to be sigmoidal. Hornik (1993) provides results that permit

even more general activation functions. Moreover, a feedforward network with only one hidden layer suffices for such approximation property. More hidden layers may be helpful in certain applications but are not necessary for functional approximation.

Barron (1993) further derived the rate of approximation in terms of mean squared error $\|f-g\|_2^2$. It was shown that three-layered feedforward networks $f_{G,q}$ with G a sigmoidal function can achieve the approximation rate of order $O(1/q)$, for which the number of parameters grows linearly with q (with the order $O(mq)$). This is in sharp contrast with other expansions, such as polynomial (with p the degree of the polynomial) and spline (with p the number of knots per coordinate), which yield suitable approximation when the number of parameters grows exponentially (with the order $O(p_m)$). Thus, it is practically difficult for such expansions to approximate well when the dimension of the input space, m, is large.

Implementation of ANNs

In practice, when the activation functions in an ANN are chosen, it remains to estimate its connection weights (unknown parameters) and to determine a proper number of hidden units. Given that the connection weights of an ANN model are unknown, this network must be properly 'trained' so as to 'learn' the unknown weights. This is why parameter estimation is referred to as network *learning* and the sample used for parameter estimation is referred to a *training sample* in the ANN literature. As the number of hidden units q determines network complexity, finding a suitable q is known as network *complexity regularization*.

Model Estimation

The network parameters can be estimated by either *online* or *offline* methods. An online learning algorithm is just a *recursive estimation* method which updates parameter estimates when new sample information becomes available. By contrast, offline learning methods are based on fixed training samples; standard econometric estimation methods are typically offline.

To ease the discussion of model estimation, we focus on the simple case that there is only one target variable y and the network function $f_{G,q}$. Generalization to the case with multiple target variables and vector-valued network functions is straightforward. Once the activation function G is chosen and the number of hidden units is given, $f_{G,q}$ is a nonlinear parametric model for the target y; the network with multiple outputs is a system of nonlinear models. If we take mean squared error as the criterion, the parameter vector of interest $\boldsymbol{\theta}^*$ thus minimizes

$$\mathbb{E}\left[y - f_{G,q}(\mathbf{x};\boldsymbol{\theta})\right]^2. \tag{6}$$

It is well known that

$$\mathbb{E}\left[y - f_{G,q}(\mathbf{x};\boldsymbol{\theta})\right]^2 = \mathbb{E}[y - E(y|\mathbf{x})]^2 + \mathbb{E}\left[\mathbb{E}(y|\mathbf{x}) - f_{G,q}(\mathbf{x};\boldsymbol{\theta})\right]^2.$$

As $\mathbb{E}(y|\mathbf{x})$ is the best L_2 predictor of y, $\boldsymbol{\theta}^*$ must also minimize the mean squared approximation error: $\mathbb{E}\left[\mathbb{E}(y|\mathbf{x}) - f_{G,q}(\mathbf{x};\boldsymbol{\theta})\right]^2$. This shows that, among all three-layered feedforward networks with the activation function G and q hidden units, $f_{G,q}(\mathbf{x};\boldsymbol{\theta}^*)$ provides the best approximation to the conditional mean function.Given a training sample of T observations, an estimator of $\boldsymbol{\theta}^*$ can be obtained by minimizing the sample counterpart of (6):

$$\frac{1}{T}\sum_{t=1}^{T}\left[y_t - f_{G,q}(\mathbf{x};\boldsymbol{\theta})\right]^2,$$

which is just the objective function of the nonlinear least squares (NLS) method. The NLS method is an offline estimation method because the size of the training sample is fixed. Under very general conditions on the data and nonlinear function, it is well known that the NLS estimator is strongly consistent for $\boldsymbol{\theta}^*$ and asymptotically normally distributed (see, for example, Gallant and White 1988).

In many ANN applications (for example, signal processing and language learning), the training sample is not fixed but constantly expands with new data. In such cases, offline estimation

may not be feasible, but online estimation methods, which update the parameter estimates based solely on the newly available data, are computationally more tractable. Moreover, online estimation methods can be interpreted as 'adaptive learning' by biological neural systems. It should be emphasized that when there is only a given sample, as in most empirical studies in economics, recursive estimation is *not* to be preferred because it is, in general, statistically less efficient than the NLS method in finite samples.

Note that the parameter of interest θ^* is the zero of the first order condition of (6):

$$\mathbb{E}\left[\nabla f_{G,q}(\mathbf{x};\boldsymbol{\theta})\left(y - f_{G,q}(\mathbf{x};\boldsymbol{\theta})\right)\right]^2 = 0,$$

where $\nabla f_{G,q}(\mathbf{x};\boldsymbol{\theta})$ is the (column) gradient vector of $f_{G,q}$ with respect to $\boldsymbol{\theta}$. To estimate $\boldsymbol{\theta}^*$, a recursive algorithm proposed by Rumelhart et al. (1986a) is

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \eta_t \nabla f_{G,q}\left(\mathbf{x}\widehat{\boldsymbol{\theta}}\right)\left[y_t - f_{G,q}\left(\mathbf{x}\widehat{\boldsymbol{\theta}}\right)\right], \quad (7)$$

where $\eta_t > 0$ is a parameter that re-scales the adjustment term in the square bracket. It can be seen from (7) that the adjustment term is determined by the gradient descent direction and the error between the target and network output: $y_t - f_{G,q}\left(\mathbf{x}_t;\widehat{\boldsymbol{\theta}}_t\right)$, and it requires only the information at time t, that is, y_t, x_t, and the estimate $\widehat{\boldsymbol{\theta}}_t$. (The algorithm (7) is analogous to the numerical steepest-descent algorithm. However, (7) utilizes only the information at time t, whereas numerical optimization algorithms are computed using all the information in a given sample and hence are offline methods.)

The algorithm (7) is known as the *error back-propagation* (or simply back-propagation) algorithm in the ANN literature, because the error signal $\left[y_t - f_{G,q}\left(\mathbf{x}_t;\widehat{\boldsymbol{\theta}}_t\right)\right]$ is propagated back through the network to determine the change of each weight. The underlying idea of this algorithm can be traced back to the classical *stochastic approximation* method introduced in Robbins and Monro (1951). White (1989) established consistency and asymptotic normality of $\widehat{\boldsymbol{\theta}}_t$ in (7). Note that the parameter η_t in the algorithm is known as a *learning rate*. For consistency of $\widehat{\boldsymbol{\theta}}_t$ it is required that η_t satisfies $\sum_{t=1}^{\infty}\eta_t = \infty$ and $\sum_{t=1}^{\infty}\eta_t^2 < \infty$, for example, $\eta_t = 1/t$. The former condition ensures that the updating process may last indefinitely, whereas the latter implies $\eta_t^2 \to 0$ so that the adjustment in the parameter estimates can be made arbitrarily small. (In many applications of ANN, the learning rate is often set to a constant η_o; the resulting estimate $\widehat{\boldsymbol{\theta}}_t$ loses consistency in this case. Kuan and Hornik (1991) established a convergence result based on small $-\eta_o$ asymptotics.)

Instead of the gradient descent direction, it is natural to construct a recursive algorithm with a Newton search direction. Kuan and White (1994) proposed the following algorithm:

$$\begin{aligned}\widehat{\mathbf{H}}_{t+1} &= \widehat{\mathbf{H}}_t + \eta_t\left[\nabla f_{G,q}\left(\mathbf{x}_t;\widehat{\boldsymbol{\theta}}_t\right)\nabla f_{G,q}\left(\mathbf{x}\widehat{\boldsymbol{\theta}}\right)' - \widehat{\mathbf{H}}_t\right],\\ \widehat{\boldsymbol{\theta}}_{t+1} &= \widehat{\boldsymbol{\theta}}_t + \eta_t\widehat{\mathbf{H}}_{t+1}^{-1}\nabla f_{G,q}\left(\mathbf{x}_t;\widehat{\boldsymbol{\theta}}_t\right)\\ &\quad\left[y_t - f_{G,q}\left(\mathbf{x}_t;\widehat{\boldsymbol{\theta}}_t\right)\right],\end{aligned} \quad (8)$$

Where $\widehat{\mathbf{H}}_{t+1}$ characterizes a Newton direction and is recursively updated via the first equation. Kuan and White (1994) showed that $\widehat{\boldsymbol{\theta}}_t$ in (8) is $\sqrt{1}$-consistent, statistically more efficient than $\widehat{\boldsymbol{\theta}}_t$ in (7), and asymptotically equivalent to the NLS estimator. The algorithm (8) may be implemented in different ways; for example, there is an algorithm that is algebraically equivalent to (8) but does not involve matrix inversion. See Kuan and White (1994) for more discussions on the implementation of the Newton algorithms.

On the other hand, estimating recurrent networks is more cumbersome. From (4) and (5) we can see that recurrent network functions depend on $\boldsymbol{\theta}$ directly and also indirectly through the presence of internal feedbacks (that is, lagged output and lagged hidden-unit activations). The indirect dependence on parameters must be taken into account in calculating the derivatives with respect to $\boldsymbol{\theta}$. Thus, NLS optimization algorithms that require analytic derivatives are difficult to implement. Kuan et al. (1994) proposed the *dynamic back-propagation* algorithm for recurrent networks,

which is analogous to (7) but involves more updating equations. Kuan (1995) further proposed a Newton algorithm for recurrent networks, analogous to (8), and showed that it is $\sqrt{1}$-consistent and statistically more efficient than the dynamic backpropagation algorithm. We omit the details of these algorithms; see Kuan and Liu (1995) for an application of these estimation methods for both feedforward and recurrent networks.

Note that the NLS method and recursive algorithms all require computing the derivatives of the network function. Thus, a smooth and differentiable activation function, as the examples given in section "Choices of activation function", are quite convenient for network parameter estimation. Finally, given that ANN models are highly nonlinear, it is likely that there exist multiple optima in the objective function. There is, however, no guarantee that the NLS method and the recursive estimation methods discussed above will deliver the global optimum. This is a serious problem because the dimension of the parameter space is typically large. Unfortunately, a convenient and effective method for finding the global optimum in ANN estimation is not yet available.

Model Complexity Regularization

Section "ANN as an universal approximator" shows that a network model $f_{G,q}$ can approximate unknown function when the number of hidden units, q, is sufficiently large. When there is a fixed training sample, a complex network with a very large q may over fit the data. Thus, there is a trade-off between approximation capability and over-fitting in implementing ANN models.

An easy approach to regularizing the network complexity is to apply a model selection criterion, such as Schwarz (Bayesian) information criterion (BIC), to the network models with various q. (Alternatively, one may consider testing whether some hidden units may be dropped from the model. This amounts to testing, say, $\beta_h = 0$ for some h. Unfortunately, the parameters in that hidden-unit activation function ($\gamma_{h,0}$ and γ_h) are not identified under this null hypothesis. It is well known that, when there are unidentified nuisance parameters, standard econometric tests are not applicable.) As is well known, BIC consists of two terms: one is based on model fitness, and the other penalizes model complexity. Hence, it is suitable for regularizing network complexity; see also Barron (1991). A different criterion introduced in Rissanen (1986, 1987) is *predictive stochastic complexity* (PSC) which is just an average of squared prediction errors:

$$\mathrm{PSC} = \frac{1}{T-k}\sum_{t=k=1}^{T}\left[y_t - f_{G,q}\left(\mathbf{x}_t;\widehat{\boldsymbol{\theta}}_t\right)\right]^2,$$

where $\widehat{\boldsymbol{\theta}}_t$ is the predicted parameter estimate based on the sample information up to time $t-1$, and k is the total number of parameters in the network. Given the number of inputs, the network with the smallest BIC or PSC gives the desired number of hidden units $\boldsymbol{q}^*$. Rissanen showed that both BIC and PSC can be interpreted as the criteria for 'minimum description length' in the sense that they determine the shortest code length (asymptotically) that is needed to encode a sequence of numbers. In other words, these criteria lead to the least complex model that still captures the key information in data. Swanson and White (1997) showed that a network selected by BIC need not perform well in out-of-sample forecasting, however.

Clearly, PSC requires estimating the parameters at each t. It would be computationally demanding if the NLS method is to be used, even for a moderate sample. For simplicity, Kuan and Liu (1995) suggested a two-step procedure for implementing ANN models. In the first step, one estimates the network models and computes the resulting PSCs using the recursive Newton algorithm, which is asymptotically equivalent to the NLS method. When a suitable network structure is determined, the Newton parameter estimates can be used as initial values for NLS estimation in the second step. This approach thus maintains a balance between computational cost and estimator efficiency.

Concluding Remarks

In this article, we introduce ANN model specifications, their approximation properties, and the

methods for model implementation from an econometric perspective. It should be emphasized that ANN is neither a magical econometric tool nor a 'black box' that can solve *any* difficult problems in econometrics. As discussed above, a major advantage of ANN is its universal approximation property, a property shared by other nonparametric approximators. Yet, compared with parametric econometric models, a simple ANN need not perform better, and a more complex ANN (with a large number of hidden units) is more difficult to implement properly and cannot be applied when there is only a small data-set. Therefore, empirical applications of ANN models must be exercised with care.

See Also

- ► Non-parametric Structural Models
- ► Stochastic Adaptive Dynamics

Acknowledgment *I would like to express my sincere gratitude to Steven Durlauf for his patience and constructive comments on early drafts of this article. I also thank Shih-Hsun Hsu and Yu-Lieh Huang for very helpful suggestions. The remaining errors are all mine.*

Bibliography

Barron, A. 1991. Complexity regularization with application to artificial neural networks. In *Nonparametric functional estimation and related topics*, ed. G. Roussas. Dordrecht: Kluwer Academic Publishers.

Barron, A. 1993. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* 39: 930–945.

Cybenko, G. 1989. Approximation by superpositions of a sigmoid function. *Mathematics of Control, Signals, and Systems* 2: 303–314.

Elman, J. 1990. Finding structure in time. *Cognitive Science* 14: 179–211.

Funahashi, K. 1989. On the approximate realization of continuous mappings by neural networks. *Neural Networks* 2: 183–192.

Gallant, A., and H. White. 1988. *A unified theory of estimation and inference for nonlinear dynamic models*. Oxford: Basil Blackwell.

Hornik, K. 1991. Approximation capabilities of multilayer feedforward nets. *Neural Networks* 4: 231–242.

Hornik, K. 1993. Some new results on neural network approximation. *Neural Networks* 6: 1069–1072.

Hornik, K., M. Stinchcombe, and H. White. 1989. Multi-layer feedforward networks are universal approximators. *Neural Networks* 2: 359–366.

Hornik, K., M. Stinchcombe, and H. White. 1990. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks* 3: 551–560.

Jordan, M. 1986. *Serial order: A parallel distributed processing approach*, Report No. 8604. La Jolla: Institute for Cognitive Science, University of California San Diego.

Kuan, C.-M. 1995. A recurrent Newton algorithm and its convergence properties. *IEEE Transactions on Neural Networks* 6: 779–783.

Kuan, C.-M., and K. Hornik. 1991. Convergence of learning algorithms with constant learning rates. *IEEE Transactions on Neural Networks* 2: 484–489.

Kuan, C.-M., and T. Liu. 1995. Forecasting exchange rates using feedforward and recurrent neural networks. *Journal of Applied Econometric* 10: 347–364.

Kuan, C.-M., and H. White. 1994. Artificial neural networks: An econometric perspective (with discussions). *Econometric Reviews* 13(1–91): 139–143.

Kuan, C.-M., K. Hornik, and H. White. 1994. A convergence result for learning in recurrent neural networks. *Neural Computation* 6: 420–440.

McClelland, J.,D. Rumelhart, and the PDP Research Group. 1986. *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 2. Cambridge, MA: MIT Press.

McCulloch, W., and W. Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5: 115–133.

Rissanen, J. 1986. Stochastic complexity and modeling. *Annals of Statistics* 14: 1080–1100.

Rissanen, J. 1987. Stochastic complexity (with discussions). *Journal of the Royal Statistical Society, B* 49(223–39): 252–265.

Robbins, H., and S. Monro. 1951. A stochastic approximation method. *Annals of Mathematical Statistics* 22: 400–407.

Rosenblatt, F. 1958. The perception: A probabilistic model for information storage and organization in the brain. *Psychological Reviews* 62: 386–408.

Rumelhart, D., G. Hinton, and R. Williams. 1986a. Learning internal representation by error propagation. In *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1, ed. D. Rumelhart, J. McClelland and the PDP Research Group. Cambridge, MA: MIT Press.

Rumelhart, D., J. McClelland, and the PDP Research Group, eds. 1986b. *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1. Cambridge, MA: MIT Press.

Stinchcombe, M., and H. White. 1998. Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory* 14: 295–325.

Swanson, N., and H. White. 1997. A model selection approach to real-time macroeconomic forecasting

using linear models and artificial neural networks. *Review of Economics and Statistics* 79: 540–550.
White, H. 1988. *Economic prediction using neural networks: The case of IBM stock prices*.In: Proceedings of the second annual IEEE conference on neural networks II. New York: IEEE Press.
White, H. 1989. Some asymptotic results for learning in single hidden layer feedforward network models. *Journal of the American Statistical Association* 84: 1003–1013.

Artificial Regressions

James G. MacKinnon

Abstract

An artificial regression is a linear regression that is associated with some other econometric model, which is usually nonlinear. It can be used for a variety of purposes, in particular computing covariance matrices and calculating test statistics. The best-known artificial regression is the *Gauss–Newton regression*, whose key properties are shared by all artificial regressions. The chief advantage of artificial regressions is conceptual: because econometricians are very familiar with linear regression models, using them for computation reduces the chance of errors and makes the results easier to comprehend intuitively.

Keywords

Artificial regressions; Binary response model regression; Bootstrap; Double-length artificial regression; Efficient score tests; Gauss–Newton regression; Generalized method of moments; Heteroskedasticity; Heteroskedasticity-consistent covariance matrices; Instrumental variables; Lagrange multiplier tests; Multivariate nonlinear regression models; Non-nested hypotheses; Outer product of the gradient regression; RESET test; Score tests; Specification

JEL Classifications

C1

An *artificial regression* is a linear regression that is associated with some other econometric model, which is usually, but not always, nonlinear. It can be used for a variety of purposes, in particular, computing covariance matrices and calculating test statistics. The best-known artificial regression is the *Gauss–Newton regression* (GNR), which is discussed in the next section. All artificial regressions share the key properties of the GNR.

The Gauss–Newton Regression

A univariate nonlinear regression model may be written as

$$y_t = x_t(\beta) + u_t, \;\; u_t \sim \text{IID}\left(0, \;\; \sigma^2\right), t = 1, \ldots, n, \tag{1}$$

where y_t is the tth observation on the dependent variable, and β is a k-vector of parameters to be estimated. Here the scalar function $x_t(\beta)$ is a nonlinear regression function which may depend on exogenous and/or predetermined variables. The model (1) may also be written using vector notation as

$$\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}, \;\; \mathrm{u} \sim \text{IID}\left(0, \;\; \sigma^2\mathbf{I}\right), \tag{2}$$

where $\boldsymbol{y}$ is an n-vector with typical element y_t, $\mathrm{x}(\boldsymbol{\beta})$ is an n-vector with typical element $x_t(\boldsymbol{\beta})$, and $\mathbf{I}$ is an $n \times n$ identity matrix.

The Gauss–Newton regression that corresponds to (2) is

$$\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}) = \boldsymbol{X}(\boldsymbol{\beta})\boldsymbol{b} + \;\; \text{residuals}, \tag{3}$$

where $\boldsymbol{b}$ is an n-vector of regression coefficients, and the matrix $\boldsymbol{X}(\boldsymbol{\beta})$ is $n \times k$ with tith element the derivative of $x_t(\boldsymbol{\beta})$ with respect to β_i, the ith component of $\boldsymbol{\beta}$. The regressand here is a vector of residuals, and the regressors are matrices of derivatives. When regression (3) is evaluated at the least-squares estimates $\hat{\boldsymbol{\beta}}$, it becomes

$$\hat{\boldsymbol{u}} \equiv \boldsymbol{y} - \hat{\boldsymbol{x}} = \hat{\boldsymbol{X}}\mathrm{b} + \;\; \text{residuals}, \tag{4}$$

where $\hat{\boldsymbol{x}} \equiv \boldsymbol{x}\left(\hat{\boldsymbol{\beta}}\right)$ and $\hat{\boldsymbol{X}} \equiv \boldsymbol{X}\left(\hat{\boldsymbol{\beta}}\right)$. Since the regressand of this artificial regression must be orthogonal to all the regressors, running the GNR (4) is an easy way to check that the NLS estimates actually satisfy the first-order conditions.

The usual OLS covariance matrix for $\hat{\boldsymbol{b}}$ from regression (4) is

$$s^2\left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1}, \quad \text{where } s^2 = \frac{1}{n-k}(\boldsymbol{y}-\hat{\boldsymbol{x}})'(\boldsymbol{y}-\hat{\boldsymbol{x}}). \tag{5}$$

This is also the usual estimator of the covariance matrix of the NLS estimator $\hat{\boldsymbol{\beta}}$ under the assumption that the errors are IID. If that assumption were relaxed to allow for heteroskedasticity of unknown form, then (5) would be replaced by a heteroskedasticity-consistent covariance matrix (HCCME) of the form

$$\left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1}\hat{\boldsymbol{X}}'\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{X}}\left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1}, \tag{6}$$

where $\hat{\boldsymbol{\Omega}}$ is an $n \times n$ diagonal matrix with squared residuals, probably rescaled, on the principal diagonal. The matrix (6) is precisely what a regression package would give if we ran the GNR (4) and requested an HCCME. Similar results hold if we relax the independence assumption and use a HAC estimator. In every case, a standard estimator of the covariance matrix of $\hat{\boldsymbol{b}}$ from the artificial regression (4) is also perfectly valid for the NLS estimates $\hat{\boldsymbol{\beta}}$.

If we evaluate the GNR (3) at a vector of restricted estimates $\tilde{\boldsymbol{\beta}}$, we can use the resulting artificial regression to test the restrictions. For simplicity, assume that $\tilde{\boldsymbol{\beta}} = \left[\tilde{\boldsymbol{\beta}}_1' \; 0'\right]'$, where $\boldsymbol{\beta}_1$ is a k_1-vector and $\boldsymbol{\beta}_2$, which is equal to $\mathbf{0}$ under the null hypothesis, is a k_2–vector. In this case, the GNR becomes

$$\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{X}}_1\boldsymbol{b}_1 + \tilde{\boldsymbol{X}}_2\boldsymbol{b}_2 + \text{ residuals}. \tag{7}$$

The ordinary F statistic for $\boldsymbol{b}_2 = 0$ is asymptotically valid as a test for $\boldsymbol{\beta}_2 = 0$, and it is asymptotically equal, under the null hypothesis, to the F statistic for $\boldsymbol{\beta}_2 = 0$ in the nonlinear regression (1). Of course, when $\boldsymbol{X}_2$ has just one column, the t statistic for the scalar b_2 to equal zero is also asymptotically valid. Yet another test statistic that is frequently used is n times the uncentred R^2 from regression (7), which is asymptotically distributed as $\chi^2(k_2)$ under the null hypothesis.

The GNR (3) can also be used as part of a quasi-Newton minimization procedure if it is evaluated at any vector, say $\boldsymbol{\beta}_{(j)}$, where j denotes the jth step of an iterative procedure. In fact, this is where the name of the GNR came from. It is not hard to show that the vector

$$\boldsymbol{b}_{(j)} \equiv \left(\boldsymbol{X}_{(j)}'\boldsymbol{X}_{(j)}\right)^{-1}\boldsymbol{X}_{(j)}\left(\boldsymbol{y}-\boldsymbol{x}_{(j)}\right),$$

where the notation should be obvious, is asymptotically equivalent to the vector that defines a Newton step starting at $\boldsymbol{\beta}_{(j)}$. The vector $\boldsymbol{b}_{(j)}$ is asymptotically equivalent to what we would get by postmultiplying minus the inverse of the Hessian of the sum of squared residuals function by the gradient. Because of this, the GNR has the same *one-step property* as Newton's method itself. If we evaluate (3) at any consistent estimator, say $\ddot{\boldsymbol{\beta}}$, then the one-step estimator $\boldsymbol{\beta}' = \ddot{\boldsymbol{\beta}} + \ddot{\boldsymbol{b}}$ is asymptotically equivalent to the NLS estimator $\hat{\boldsymbol{\beta}}$.

For more detailed treatments of the Gauss–Newton regression, see MacKinnon (1992) and Davidson and MacKinnon (2001, 2004).

Properties of Artificial Regressions

A very general class of artificial regressions can be written as

$$\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{R}(\boldsymbol{\theta})\boldsymbol{b} + \text{ residuals}, \tag{8}$$

where $\boldsymbol{\theta}$ is a parameter vector of length k, $\mathrm{r}(\boldsymbol{\theta})$ is a vector of length an integer multiple of the sample size n, and $\boldsymbol{R}(\boldsymbol{\theta})$ is a matrix with k columns and as many rows as $\boldsymbol{r}(\boldsymbol{\theta})$. In order to qualify as an artificial regression, the linear regression (8) must satisfy three key properties.

1. The regressand $\boldsymbol{r}\left(\hat{\boldsymbol{\theta}}\right)$ is orthogonal to every column of the matrix of regressors $\boldsymbol{R}\left(\hat{\boldsymbol{\theta}}\right)$, where $\hat{\boldsymbol{\theta}}$ denotes a vector of unrestricted estimates. That is,

$$\boldsymbol{R}'\left(\hat{\boldsymbol{\theta}}\right)\boldsymbol{r}\left(\hat{\boldsymbol{\theta}}\right) = 0. \tag{9}$$

2. The asymptotic covariance matrix of $n^{1/2}\left(\hat{\theta} - \theta_0\right)$ is given either by

$$\operatorname*{plim}_{n\to\infty} \left(n^{-1}\boldsymbol{R}'\left(\hat{\boldsymbol{\theta}}\right)\boldsymbol{R}\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1}, \text{ or by} \tag{10}$$

$$\operatorname*{plim}_{n\to\infty} s^2\left(\boldsymbol{n}^{-1}\boldsymbol{R}'\left(\hat{\boldsymbol{\theta}}\right)\boldsymbol{R}\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1}, \tag{11}$$

where s^2 is the OLS estimate of the error variance obtained by running regression (8) with $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Of course, this is also tme if $\hat{\boldsymbol{\theta}}$ is replaced by any other consistent estimator of $\boldsymbol{\theta}$.

3. If $\ddot{\boldsymbol{\theta}}$ denotes a consistent estimator, and $\ddot{\boldsymbol{b}}$ denotes the vector of estimates obtained by mnning regression (8) evaluated at $\ddot{\boldsymbol{\theta}}$, then

$$\begin{aligned}\operatorname*{plim}_{n\to\infty} \quad & n^{1/2}\left(\ddot{\boldsymbol{\theta}} + \ddot{\boldsymbol{b}} - \boldsymbol{\theta}_0\right) \\ = \operatorname*{plim}_{n\to\infty} \quad & n^{1/2}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right).\end{aligned} \tag{12}$$

This is the one-step property, which holds because the vector $\ddot{\boldsymbol{b}}$ is asymptotically equivalent to a single Newton step.

There exist many artificial regressions that take the form of (8) and satisfy conditions 1, 2, and 3. Some of these will be discussed in the next section. We have seen that the GNR satisfies these conditions and that its asymptotic covariance matrix is given by (11).

The most widespread use of artificial regressions is for specification testing. Of course, any artificial regression can be used to test restrictions on the model to which it corresponds. We simply evaluate the artificial regression for the unrestricted model at the restricted estimates, as in (7). However, in many cases, we can also use artificial regressions to test model specification without explicitly specifying an alternative. Consider the artificial regression

$$\boldsymbol{r}\left(\hat{\boldsymbol{\theta}}\right) = \boldsymbol{R}\left(\hat{\boldsymbol{\theta}}\right)\boldsymbol{b} + Z\left(\hat{\boldsymbol{\theta}}\right)c + \text{ residuals}, \tag{13}$$

which is evaluated at unrestricted estimates $\hat{\boldsymbol{\theta}}$. Here $\boldsymbol{Z}(\boldsymbol{\theta})$ is a matrix with r columns, each of which is supposed to be asymptotically uncorrelated with $\boldsymbol{r}(\boldsymbol{\theta})$, that has certain other properties which ensure that standard test statistics for $\boldsymbol{c} = 0$ are asymptotically valid. In effect, regression (13) must have the same properties as if it corresponded to an unrestricted model. See Davidson and MacKinnon (2001, 2004) for details.

When the artificial regression (13) is a GNR, $\boldsymbol{r}\left(\hat{\boldsymbol{\theta}}\right) = \hat{\boldsymbol{u}}$ and $\boldsymbol{R}\left(\hat{\boldsymbol{\theta}}\right) = \hat{\boldsymbol{X}}$. Such a GNR can be used to implement a number of well-known specification tests, including the following ones.

- If we let $\boldsymbol{Z}\left(\hat{\boldsymbol{\theta}}\right)$ be a vector of squared fitted values, then the t statistic for the coefficient on the test regressor to be zero can be used to perform one version of the well-known RESET test (Ramsey, 1969).
- If we let $\boldsymbol{Z}\left(\hat{\boldsymbol{\theta}}\right)$ be an $n \times p$ matrix containing the residuals lagged once through p times, either the F statistic for $\boldsymbol{c} = 0$ or n times the uncentred R^2 can be used to perform a standard test for pth order serial correlation (Godfrey, 1978).
- If we let $\boldsymbol{Z}\left(\hat{\boldsymbol{\theta}}\right)$ be the vector $\hat{\boldsymbol{w}} - \hat{\boldsymbol{x}}$, where $\hat{\boldsymbol{w}}$ denotes the fitted values from a non-nested alternative model, then the t statistic on the test regressor can be used to perform a non-nested hypothesis test, namely, the P test proposed by Davidson and MacKinnon (1981).

Like all asymptotic tests, the three tests just described may not have good finite-sample properties. This is particularly true for the P test and other non-nested hypothesis tests. Finite-sample

properties can often be greatly improved by bootstrapping, which is quite easy to do in these cases. For a recent survey of bootstrap methods in econometrics, see Davidson and MacKinnon (2006).

More Artificial Regressions

A great many artificial regressions have been proposed over the years, far more than there is space to discuss here. Some of them apply to very broad classes of econometric models, and others to quite narrow ones.

One of the most widely applicable and commonly used artificial regressions is the *outer product of the gradient* (OPG) regression. It applies to every model for which the log-likelihood function can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(\boldsymbol{\theta}), \tag{14}$$

where ℓ_t is the contribution to the log-likelihood made by the tth observation, and $\boldsymbol{\theta}$ is a k-vector of parameters. The $n \times k$ matrix of contributions to the gradient, $\boldsymbol{G}(\boldsymbol{\theta})$, has typical element

$$\boldsymbol{G}_{ti}(\boldsymbol{\theta}) \equiv \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \theta_i}. \tag{15}$$

Summing the elements of the ith column of this matrix yields the ith element of the gradient. The OPG regression is

$$\iota = \boldsymbol{G}(\boldsymbol{\theta}) + \mathit{residuals}, \tag{16}$$

where ι is an n-vector of ones.

It is easy to see that the OPG regression satisfies condition 1, since the inner product of ι and $\boldsymbol{G}(\boldsymbol{\theta})$ is just the gradient, which must be zero when evaluated at the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$. That it satisfies condition 2 follows from the fact that the plim of the matrix $n^{-1}\boldsymbol{G}'(\boldsymbol{\theta})\boldsymbol{G}(\boldsymbol{\theta})$ is the information matrix, which implies that the asymptotic covariance matrix is give by (10). The OPG regression also satisfies condition 3, and it is therefore a valid artificial regression.

Because it applies to such a broad class of models, the OPG regression is easy to use in a wide variety of contexts. This includes information matrix tests (Chesher 1983; Lancaster 1984) and conditional moment tests (Newey 1985), both of which may be thought of as special cases of regression (13). However, because $n^{-1}\boldsymbol{G}'\left(\hat{\boldsymbol{\theta}}\right)\boldsymbol{G}\left(\hat{\boldsymbol{\theta}}\right)$ tends to be an inefficient estimator of the information matrix, tests based on the OPG regression often have poor finite-properties, iterative procedures based on it may converge slowly, and covariance matrix estimates may be poor. Davidson and MacKinnon (1992) contains some simulation results which show just how poor the finite-properties of tests based on the OPG regression can be. However, these properties can often be improved dramatically by bootstrapping.

Another artificial regression that applies to a fairly general class of models estimated by maximum likelihood is the *double-length artificial regression* (DLR), proposed by Davidson and MacKinnon (1984). The class of models to which it applies may be written as

$$f_t(y_t, \ \boldsymbol{\theta}) = \varepsilon_t, \ \ t = 1, \ldots, n, \ \ \varepsilon_t \sim \mathrm{NID}(0, 1), \tag{17}$$

where $f_t(\cdot)$ is a smooth function that depends on the random variable y_t, on a k-vector of parameters $\boldsymbol{\theta}$, and, implicitly, on exogenous and/or predetermined variables. This class of models is much more general than may be apparent at first. It includes both univariate and multivariate linear and nonlinear regression models, as well as models that involve transformations of the dependent variable. The main restrictions are that the dependent variable(s) must be continuous and that the distribution(s) of the error terms must be known.

As its name suggests, the DLR has $2n$ observations. It can be written as

$$\begin{bmatrix} \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{\theta}) \\ \iota \end{bmatrix} = \begin{bmatrix} -\boldsymbol{F}(\boldsymbol{y}, \boldsymbol{\theta}) \\ \boldsymbol{K}(\boldsymbol{y}, \boldsymbol{\theta}) \end{bmatrix} \boldsymbol{b} + \text{residuals}. \tag{18}$$

Here $\boldsymbol{f}(\boldsymbol{y}, \ \theta)$ is an n-vector with typical element $f_t(y_t, \ \theta)$, ι is an n-vector of ones, $\boldsymbol{F}(\boldsymbol{y}, \ \theta)$ is

an $n \times k$ matrix with typical element $\partial f_t(y_t, \theta)/\partial\theta_i$, and $\boldsymbol{K}(\boldsymbol{y}, \theta)$ is an $n \times k$ matrix with typical element $\partial k_t(y_t, \theta)/\partial\theta_i$, where

$$k_t(y_t, \boldsymbol{\theta}) \equiv \log\left|\frac{\partial f_t(y_t, \boldsymbol{\theta})}{\partial y_t}\right|$$

is a Jacobian term that appears in the log-likelihood function for the model (17). The information matrix associated with the DLR (18) has the form

$$\frac{1}{n}(\boldsymbol{F}'(\boldsymbol{\theta}) + \boldsymbol{F}(\boldsymbol{\theta}) + \boldsymbol{K}'(\boldsymbol{\theta})\boldsymbol{K}(\boldsymbol{\theta})). \qquad (19)$$

In most cases, this is a much more efficient estimator than the one associated with the OPG regression. As a result, inferences based on the DLR are generally more reliable than inferences based on the OPG regression. See, for example, Davidson and MacKinnon (1992). The DLR is not the only artificial regression for which the number of 'observations' is a multiple of the actual number. For other examples, see Orme (1995).

Ideally, an information matrix estimator should depend on the data only through estimates of the parameters. A Lagrange multiplier, or score, test based on such an estimator is often called an *efficient score test*. Because (19) often does not satisfy this condition, using the DLR generally does not yield efficient score tests. In contrast, at least for models with no lagged dependent variables, the GNR does yield efficient score tests, as do several other artificial regressions.

A number of somewhat specialized artificial regressions can be obtained as modified versions of the Gauss–Newton regression. These include two different forms of GNR that are robust to heteroskedasticity of unknown form, a variant of the GNR for models estimated by instrumental variables, a variant of the GNR for models estimated by the generalized method of moments, a variant of the GNR for multivariate nonlinear regression models, and the binary response model regression (BRMR), which applies to models like the logit and probit model. See Davidson and MacKinnon (2001, 2004) for detailed discussions and references.

Of course, any quantity that can be computed using an artificial regression can also be computed directly by using a matrix language. Why then use artificial regressions for computation? This is, to some extent, simply a matter of taste. One potential advantage is that most statistics packages perform least squares regressions efficiently and accurately. In my view, however, the chief advantage of artificial regressions is conceptual. Because econometricians are very familiar with linear regression models, using them for computation reduces the chance of errors and makes the results easier to comprehend intuitively.

See Also

- ▶ Non-nested Hypotheses
- ▶ Serial Correlation and Serial Dependence
- ▶ Testing

Bibliography

Chesher, A. 1983. The information matrix test: Simplified calculation via a score test interpretation. *Economics Letters* 13(1): 45–48.

Davidson, R., and J.G. MacKinnon. 1981. Several tests for model specification in the presence of alternative hypotheses. *Econometrica* 49: 781–793.

Davidson, R., and J.G. MacKinnon. 1984. Model specification tests based on artificial linear regressions. *International Economic Review* 25: 485–502.

Davidson, R., and J.G. MacKinnon. 1992. A new form of the information matrix test. *Econometrica* 60: 145–157.

Davidson, R., and J.G. MacKinnon. 2001. Artificial regressions. In *Companion to theoretical econometrics*, ed. B. Baltagi. Oxford: Blackwell.

Davidson, R., and J.G. MacKinnon. 2004. *Econometric theory and methods*. New York: Oxford University Press.

Davidson, R., and J.G. MacKinnon. 2006. Bootstrap methods in econometrics. In *Palgrave handbooks of econometrics. volume 1: Econometric theory*, ed. T.C. Mills and K.D. Patterson. Basingstoke: Palgrave Macmillan.

Godfrey, L.G. 1978. Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica* 46: 1303–1310.

Lancaster, T. 1984. The covariance matrix of the information matrix test. *Econometrica* 52: 1051–1053.

MacKinnon, J.G. 1992. Model specification tests and artificial regressions. *Journal of Economic Literature* 30: 102–146.

Newey, W.K. 1985. Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53: 1047–1070.

Orme, C.D. 1995. On the use of artificial regressions in certain microeconometric models. *Econometrica* 11: 290–305.

Ramsey, J.B. 1969. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society, Series B* 31: 350–371.

Asgill, John (1659–1738)

H. R. Tedder

Asgill was born at Hanley Castle, Worcestershire, 1659, called to the English bar 1692, expelled from the Irish House of Commons 1703, and from the British Parliament 1707, for an eccentric pamphlet contending that man could be translated to heaven without dying. He left this world in the ordinary way, 1738. He wrote the following economic works:

Several Assertions proved in order to create another Species of Money than Gold and Silver, London, 1696 (based on the theory 'man deals in nothing but earth'; a contemporary pamphlet asserts that it is plagiarized from J. Briscoe's *Discourse on the Late Funds*, 1694).

Essay on a Registry for Titles of Land London, 1698 (4th edn. 1758). *A Collection of Tracts*, London, 1715, 8 parts, *Abstract of the Publick Funds Granted and Continued to the Crown since 1 W. & M.*, London, 1715, (Reprinted in Somers's Tracts, 1815, xiii. p. 730–741).

Reprinted from *Palgrave's Dictionary of Political Economy.*

Ashley, William James (1860–1927)

O. Kurer

Sir William Ashley graduated from Oxford, remained there for several years as a tutor, and was elected fellow of Lincoln College in 1885. From 1888 to 1892 he was Professor of Political Economy and Constitutional History at the University of Toronto, and in 1892 was appointed to the world's first chair in economic history, at Harvard. In 1901 he became a professor in Birmingham, where he helped to organize the first university school of commerce in Britain.

Ashley was drawn towards historical economics by Toynbee and Cliffee Leslie. From early on, his anti-theoretical, ethical and empirical bent drew him into conflict with Marshall, who feared that Ashley's teaching was detrimental to the success of his own analytical and theoretical programme. Later Ashley became closely associated with the German Historical School. As an evolutionist he accepted their research programme of searching for stages of development, but admitted later in his life that its result had not lived up to the early hopes.

He was an important and successful pioneer in economic history, both by his own contributions where he helped to lay the foundation of modern economic history, and by his efforts to establish it as an acknowledged discipline.

Ashley insisted that economic theories closely reflected economic reality, were only true relative to time and circumstances, and held that influential old doctrines, generally believed to be erroneous, had not been without truth and value in their time.

Ashley consistently attacked laissez-faire, advocating further social legislation and the extension of state ownership. He supported trade unions since they, together with trusts regulated by the state, would limit competition, which in his view was responsible for crises and unemployment. He was an important participant in the Tariff controversy in 1903 where he came down in favour of protection.

See Also

► Historical Economics, British

Selected Works

1888–93. *An introduction to english economic history and theory*, 2 vols. London/New York: Longmans.

1900. *Surveys, historic and economic*. London/New York: Longmans.
1903. *The tariff problem.* London: P.S. King.
1914. *The economic organisation of England: an outline history.* London/New York: Longmans.

Bibliography

Ashley, A. 1932. *William James Ashley: A life*. London: King.
Clapham, J.H. 1927. Obituary: Sir William Ashley. *Economy Journal* 37: 678–683.
MacDonald, J.L. 1942. Sir William Ashley. In *Some historians of Modern Europe: In historiography by former students of the University of Chicago*, ed. B.E. Schmitt. Chicago: University of Chicago Press.
Samuels, W.J. 1977. Ashley's and Taussig's lectures on the history of economic thought at Harvard, 1896–1897. *History of Political Economy* 9: 384–411.
Semmel, B. 1957. Sir William Ashley as 'Socialist of the Chair'. *Economica, NS* 24: *343–411*.
Wood, J.C. 1983. *British economists and the empire, 1860–1914*. Beckenham: Croom Helm.

Ashton, Thomas Sutcliffe (1889–1968)

Phyllis Deane

Keywords
Ashton, T. S; Economic history

T.S. Ashton was born in Lancashire in 1889, graduated from Manchester University in 1909 and returned there in 1921 (after some years at the Universities of Sheffield and Birmingham) to teach political economy and economic history in the Faculty of Commerce. By the time he took up Eileen Power's chair of economic history at the London School of Economics in 1944, he had made a substantive and distinctive contribution to the history of the industrial revolution in three research monographs: *Iron and Steel and in the Industrial Revolution* (1924), *The Coal Industry of the Eighteenth Century* (1929, written with Joseph Sykes), and *An Eighteenth Century Industrialist: Peter Stubs of Warrington* (1939). Over the next decade this unassuming, humane, passionately non-dogmatic scholar had become the leader of a new generation of economic historians, a generation whose members had been schooled in the theories and analytical techniques of economics rather than in the thinking habits of a history faculty.

The two industrial studies and the business history published while Ashton was in Manchester were exercises in applied economics, based on detailed investigation of primary sources (including a mass of business ledgers, letters and accounts) and of a wide range of 18th-century material reflecting economic and social events, transactions and opinions. These researches gave him a formidable armoury of qualitative and quantitative data from which he set out in the 1940s explicitly to 'find answers (partial and provisional though these may be) to the questions economists ask, or should ask, of the past'.

Ashton's last three books constituted a coherent and cumulative contribution to the economic history of the first country to make the transition to modern economic growth. His highly original essay *The Industrial Revolution* (1948) appeared just when the industrialization problems of developing countries were assuming major importance on the applied economists' research agenda and became a long-running bestseller. His *Economic History of England: The Eighteenth Century* (1955), the prime example of a new genre of economic history, contained the first systematic attempt to use standard economic theory to explain long-term changes in the general level of prices and economic activity over that century, and also injected a characteristic objectivity into the perennial controversy over the standard of living of workers during the industrial revolution. In his last book, *Economic Fluctuations in England 1700–1800* (1959), he shifted his analysis of 18th-century economic change to a short-run focus. But by then only the pure theorists and the econometricians were actively interested in cyclical analysis, and Ashton was effectively distanced from both groups by his persistent concern with taking account of social as well as economic factors in economic change and by his realistically

discriminating approach to the use of either abstract concepts or statistical evidence.

Selected Works

1924. *Iron and steel in the industrial revolution*. Manchester: Manchester University Press.
1929. (With Joseph Sykes). *The coal industry of the eighteenth century*. Manchester: Manchester University Press.
1939. *An eighteenth century industrialist: Peter Stubs of Warrington*. Manchester: Manchester University Press.
1948. *The industrial revolution*. London: Oxford University Press.
1949. The standard of life of the workers of England, 1790–1830. *Journal of Economic History* 9(Suppl): 19–38.
1955. *An economic history of England: The eighteenth century*. London: Methuen.
1959. *Economic fluctuations in England, 1700–1800*. Oxford: Clarendon Press.

Asset Pricing

Thomas E. Copeland and J. Fred Weston

In the early 1950s Harry Markowitz developed a theory of portfolio selection which has resulted in a revolution in the theory of finance leading to the development of modern capital market theory (1952, 1959). He formulated a theory of investor investment selection as a problem of utility maximization under conditions of uncertainty. Markowitz discusses mainly the special case in which investors' preferences are assumed to be defined over the mean and variance of the probability distribution of single-period portfolio returns, but he also treated most issues developed more fully in the subsequent literature.

J. Tobin (1958) utilized the foundations of portfolio theory to draw implications with regard to the demand for cash balances. He also demonstrated that given the possibility of an investment in a risk-free asset as well as in a risky asset (or portfolio), an investor can construct a combined portfolio of the two assets to achieve any desired combination of risk and return. Subsequently, W. F. Sharpe, using one of the efficient methods for constructing portfolios discussed in the appendices to the Markowitz book (1959), developed what he called the 'diagonal model' in his dissertation under the direction of Markowitz, the results of which were later summarized in an article (1963). This represented another step towards general equilibrium models of asset prices developed almost simultaneously by Treynor (1965), Sharpe (1964, 1970), Lintner (1965a, b), and Mossin (1966, 1969). Important contributions were made by Fama (1971, 1976) and by Fama and Miller (1972).

These works resulted in the development of the relationship between return and risk summarized in what has been called the Security Market Line of the Capital Asset Pricing Model (CAPM).

$$E(R_j) = R_F + \left[\frac{E(R_M) - R_F}{\sigma_M^2}\right] \mathrm{COV}(R_j, R_M). \quad (1)$$

This equation says that the return required (*ex ante*) by investors on any asset is equal to the return, R_F, on a risk-free asset plus an adjustment for risk. Alternatively, the risk adjustment can be defined as the market risk premium weighted by the risk of the individual asset normalized by the variance of market returns. This latter measure has been referred to as the beta measure (β) of the risk of an individual asset or security $\left[\beta = \mathrm{COV}(R_j, R_M)/\sigma_M^2\right]$. Leading synthesis papers on the CAPM are by Jensen (1972) and Rubinstein (1973).

The CAPM model assumes that the market functions in a reasonably perfect way in the sense that: all individuals act as if they are price-takers of all relevant prices; all securities are perfectly divisible and can be sold both long and short without margin and/or escrow requirements; there are no transaction costs or taxes; and, as in nearly all useful economic theory, arbitrage opportunities are absent so that an appropriate one price law

obtains. Individuals are assumed to be risk averse, expected utility maximizers. In that differential assessment of probabilities generally explains too much, it is usual (although not necessary for all purposes) to require that probability beliefs are homogeneous (Krouse 1986). Subsequent work established that the main principles of the CAPM held up with the successive relaxation of the above assumptions (Black 1972; Brennan 1971; Lintner 1969; Mayers 1972, 1973; Merton 1973).

Roll's critique (1977) has had a major impact. His major conclusions are: (1) The only legitimate test of the CAPM is whether or not the market portfolio (which includes *all* assets) is mean-variance efficient; (2) If performance is measured relative to an index which is ex post efficient, then from the mathematics of the efficient set, no security will have abnormal performance when measured as a departure from the Security Market Line; (3) If performance is measured relative to an ex post inefficient index, then any ranking of portfolio performance is possible depending on which inefficient index has been chosen. The Roll critique does not imply that the CAPM is invalid, but that tests of the CAPM are joint tests with market efficiency and that its uses must be implemented with due care.

Three basic types of models of asset pricing have been most frequently employed. The simplest, called the *market model*, is based on the fact that returns on security j can be linearly related to returns on a 'market' portfolio, namely:

$$R_{jt} = a_j + b_j R_{Mt} + \varepsilon_{jt} \tag{2}$$

where ε_{jt} is the mean zero classical normally distributed error term. The market model assumes that the slope and intercept terms are constant over the time period during which the model is fit to the available data, a strong assumption.

The second model is the capital asset pricing theory. It requires the intercept term to be equal to the risk-free rate, or the rate of return on the minimum variance zero-beta portfolio, both of which may change over time. In its simplest form, the CAPM is written

$$R_{jt} - R_{Ft} = [R_{Mt} - R_{Ft}]\beta_{jt} + \varepsilon_{jt}. \tag{3}$$

Systematic risk, β_{jt}, is generally assumed to remain constant over the interval of estimation.

The third model is the empirical counterpart to the CAPM, referred to as the *empirical market line*

$$R_{jt} = \widehat{\gamma}_{0t} + \widehat{\gamma}_{1t}\beta_{jt} + \varepsilon_{jt}. \tag{4}$$

This formulation does not require that the intercept term equal the risk-free rate. No parameters are assumed to be constant over time. In contrast to the market model, which is a time series expression, both the intercept, $\widehat{\gamma}_{0t}$, and the slope, $\widehat{\gamma}_{1t} = (R_{Mt} - R_{Ft})$, are the estimates taken from cross-section data each time period (typically each month). The betas in Eq. 4 are (following Fama and MacBeth 1973) calculated from the market model (Eq. 2). (See Copeland and Weston, 1983, Chaps. 7 and 10).

Empirical tests of the CAPM were conducted by Miller and Scholes (1972), Fama and MacBeth (1973), and Reinganum (1981), among others. Most of the studies use monthly total returns (dividends are reinvested) on listed common stocks.

Asset pricing models have been used to measure portfolio performance by mutual funds, pension fund advisers, etc., and in residual analysis of the impact of accounting reports, stock splits, mergers, etc. Some studies have used the market model to measure the error terms or residuals-positive or negative performance. However, the generally accepted procedure is first to calculate the β's from the market line (Eq. 2). Portfolios ranked by β's provide groupings to minimize errors in the measurement of variables problem. These portfolio betas are used to develop the parameters (intercept and slope terms) in Eq. 4 which is the empirical market line used to estimate the CAPM of Eq. 3. With estimates of the γ terms, the empirical market line can then be used to calculate 'abnormal' returns or residuals from predicted security returns.

The empirical tests of CAPM typically are conducted in excess return form. The equation in this form should have an intercept term not significantly different from zero, with a slope equal to the excess market portfolio return. The

empirical tests have found an intercept term significantly above zero with a slope less than predicted. Thus the empirical securities market line is tilted clockwise implying that low beta securities earn more than the CAPM would predict and high beta securities earn less. But the main predictions of the CAPM of a positive market price for risk and a model linear in beta are supported.

The recognition that the market return alone might not explain all of the variation in the return on an asset or a portfolio gave rise to a multiple factor analysis of capital asset pricing. This more general approach formulated by Ross (1976b) was called the Arbitrage Pricing Theory (APT). Requiring only that individuals be risk averse, the APT has multiple factors and in equilibrium all assets must fall on the arbitrage pricing line. Thus the CAPM is viewed as a special case of the APT in which the return on the market portfolio is the single applicable factor.

Empirical work on the APT was performed by Gehr (1975), Roll and Ross (1980), Reinganum (1981), and Chen et al. (1984). These studies use data on equity daily rates of return for the New York and American Stock Exchange listed stocks. The initial studies establish that other factors contribute to an explanation of required returns but did not identify them. Later studies suggest that economic influences such as unexpected changes in inflation rates, default premia (measured by the difference between high- and low-grade bond yields), and the term premium in interest rates (measured by the difference between yields on short- and long-term bonds) correlate highly with the identified explanatory factors.

The CAPM and APT have provided useful conceptual frameworks for business finance applications such as capital budgeting analysis and for measurement of the cost of capital. Although the CAPM has not been perfectly validated by empirical tests, its main implications are upheld: systematic risk (beta) is a valid measure of risk, the model is linear in beta, and the tradeoff between return and risk is positive. The earliest empirical tests of the APT have shown that asset returns are explained by three or possibly four factors and have ruled out the variance of an asset's own returns as one of the factors.

See Also

- ▶ Arbitrage Pricing Theory
- ▶ Capital Asset Pricing Model
- ▶ Finance
- ▶ Intertemporal Portfolio Theory and Asset Pricing

Bibliography

Black, F. 1972. Capital market equilibrium with restricted borrowing. *Journal of Business* 45(3): 444–455.

Brennan, M.J. 1971. Capital market equilibrium with divergent borrowing and lending rates. *Journal of Financial and Quantitative Analysis* 6(5): 1197–1205.

Chen, N.F., R. Roll, and S.A. Ross. 1986. Economic forces and the stock market. *Journal of Business* 59(3): 383–403.

Copeland, T.E., and J.F. Weston. 1983. *Financial theory and corporate policy*, 2nd ed. Menlo Park: Addison-Wesley Publishing Company.

Fama, E.F. 1971. Risk, return, and equilibrium. *Journal of Political Economy* 79(1): 30–55.

Fama, E.F. 1976. *Foundations of finance*. New York: Basic Books.

Fama, E.F., and J. MacBeth. 1973. Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy* 81(3): 607–636.

Fama, E.F., and M.H. Miller. 1972. *The theory of finance*. New York: Holt, Rinehart and Winston.

Gehr Jr., A. 1975. Some tests of the arbitrage pricing theory. *Journal of the Midwest Finance Association* 7: 91–107.

Jensen, M.C. 1972. Capital markets: Theory and evidence. *Bell Journal of Economics and Management Science* 3(2): 357–398.

Krouse, C.G. 1986. *Capital markets and prices: Valuing uncertain income streams*. New York: North-Holland Press.

Lintner, J. 1965a. Security prices, risk, and maximal gains from diversification. *Journal of Finance* 20: 587–616.

Lintner, J. 1965b. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47: 13–37.

Lintner, J. 1969. The aggregation of investors' diverse judgments and preferences in purely competitive securities markets. *Journal of Financial and Quantitative Analysis* 4: 347–400.

Markowitz, H.M. 1952. Portfolio selection. *Journal of Finance* 7: 77–91.

Markowitz, H.M. 1959. *Portfolio selection: Efficient diversification of investments*. New York: Wiley.

Mayers, D. 1972. Non-marketable assets and capital market equilibrium under uncertainty. In *Studies in the*

theory of capital markets, ed. M.C. Jensen. New York: Praeger.
Mayers, D. 1973. Non-marketable assets and the determination of capital asset prices in the absence of a riskless asset. *Journal of Business* 46(2): 258–267.
Merton, R. 1973. An intertemporal capital asset pricing model. *Econometrica* 41(5): 867–887.
Miller, M., and M. Scholes. 1972. Rates of return in relation to risk: A re-examination of some recent findings. In *Studies in the theory of capital markets*, ed. M.C. Jensen, 47–78. New York: Praeger.
Mossin, J. 1966. Equilibrium in a capital asset market. *Econometrica* 34: 768–783.
Mossin, J. 1969. Security pricing and investment criteria in competitive markets. *American Economic Review* 59: 739–756.
Reinganum, M.R. 1981. The arbitrage pricing theory: Some empirical results. *Journal of Finance* 36(2): 313–322.
Roll, R. 1977. A critique of the asset pricing theory's tests: Part I. *Journal of Financial Economics* 4(2): 129–176.
Roll, R., and S. Ross. 1980. An empirical investigation of the arbitrage pricing theory. *Journal of Finance* 35(5): 1073–1103.
Ross, S.A. 1976a. Options and efficiency. *Quarterly Journal of Economics* 90(1): 75–89.
Ross, S.A. 1976b. The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13(3): 341–360.
Rubinstein, M.E. 1973. A mean-variance synthesis of corporate financial theory. *Journal of Finance* 28(1): 167–181.
Sharpe, W.F. 1963. A simplified model for portfolio analysis. *Management Science* 9: 277–293.
Sharpe, W.F. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.
Sharpe, W.F. 1970. *Portfolio theory and capital markets*. New York: McGraw-Hill.
Tobin, J. 1958. Liquidity preference as behavior toward risk. *Review of Economic Studies* 25: 65–86.
Treynor, J.L. 1965. How to rate management of investment funds. *Harvard Business Review* 43: 63–75.

Assets and Liabilities

Kenneth E. Boulding

Abstract

Assets and liabilities come into prominence when double-entry bookkeeping and the balance sheet were invented, a prerequisite for the development of complex market economies. Production is a function of the size and structure of real assets, including human capital. We derive more satisfaction from the use of assets rather than from their consumption. Like any business, the household has a balance sheet of assets and liabilities. The lack of capital accounting in government means that many of its activities have no real 'bottom line', and their value is usually assessed in noneconomic terms, sometimes resulting in catastrophic mistakes of judgement.

Keywords

Assets and liabilities; Balance sheet; Bank deposits; Bankruptcy; Capital accounting; Depreciation; Double-entry bookkeeping; Equity; Fixed capital; Great depression; Household capital; Human capital; Net national product; Working capital

JEL Classifications

M4

The concepts of assets and liabilities are very closely related. Liabilities can be regarded as negative assets. The term 'assets' is related to the French '*assez*', meaning 'enough'. It emerges as a legal concept, particularly in laws relating to bankruptcy, the question being whether in bankruptcy assets are enough to meet all the liabilities. Historically, there has been a tendency to distinguish between real, personal and equitable assets, but these distinctions are now of little importance.

In accounting, assets and liabilities come into prominence with the invention of double-entry bookkeeping and the balance sheet, a concept which seems to have originated in northern Italy at least by the 12th or 13th century. This concept was important as a prerequisite for the development of complex markets and profit-oriented economies as an improvement in the information system. Before the invention of the balance sheet it was hard for a merchant to know whether he had made any profit or not.

It is the convention of the balance sheet that assets are listed on one side and liabilities and

equity on the other side, equity being defined fundamentally as net assets; that is, assets minus liabilities, which are negative assets. Accounting practice divides both assets and liabilities into a number of categories. Assets are commonly divided into current, deferred and fixed assets. Current assets consist of cash, bank deposits, short-term notes, accrued interest, inventories of goods in process or finished goods which are expected to be sold within the accounting period, usually six months or a year. Sometimes items like repair parts are included in this category, even though their life on the shelf may be longer. Another item may be deferred assets, such as insurance, advertising payments which are paid in advance where the services have not yet been performed. Finally, there are fixed assets of a lasting nature, such as buildings and machines. There is also a category of intangible assets, like goodwill, value of patents, and so on. These tend to have a rather dubious status in accounting practice.

Liabilities have a somewhat similar categorization. Current liabilities are those which are expected to be paid off in the accounting period – wage claims, short-term loans, accounts payable, and so on. Current assets minus current liabilities is sometimes called 'working capital'. Somewhat corresponding to fixed assets are long-term loan obligations. The sum of all assets minus the sum of all liabilities is the equity or net worth. This is usually divided into paid-up capital and undistributed profits.

Every time an event happens to an organization that has a balance sheet, the items in the balance sheet change. Thus, in production, when wheat is ground into flour the stock of wheat diminishes and of flour increases. Likewise, the stock of money may diminish as wages are paid, and the product of the work is added to assets. Assets diminish as machinery and buildings depreciate. Exchanges, purchases and sales are reflected in an increase in what is acquired and a decrease in what is given up for it. When money is borrowed, cash is increased on the asset side and the debt is increased on the liability side. It is a convention of cost accounting that both exchange and production represent transfers of equal values. When something is purchased, it is valued at the amount paid for it, so that the net worth does not change. Similarly, in production, the value of what is produced is equal to what has been consumed (i.e. destroyed) in the process, whether this is the money used to pay wages, raw materials used up or depreciation.

Profit is the growth of net worth, which happens when some asset is revalued, usually at the moment of sale. If it is sold for more than the accounting cost, the difference is an increase in net worth. Before sale, the asset is valued at cost. After the sale, if it is profitable, the asset disappears from the accounts but a larger sum of money than the value of the asset is entered, and this is why the net worth increases. When profits are distributed the liquid assets are diminished and the net worth diminishes by the same amount. Interest-bearing liabilities grow at the rate of interest, which accrues. This diminishes the net worth, this being the growth of a negative asset. Interest paid, cash or some liquid asset, diminishes by the same amount as accrued interest diminishes. There is no change in the net worth. Profit is made by constant manipulation of the assets through production and exchange to increase the total value of assets at a greater rate than interest on liabilities is accruing. Debt is presumably incurred because of a belief that it will increase the total volume of assets sufficiently so that some kind of economies of scale will permit a rate of growth of the increased assets more rapid than the rate of interest on the liabilities that are incurred in order to expand the assets.

An important problem in accounting, by no means satisfactorily solved, is how to deal with inflation and deflation. In order to get a net worth or 'bottom line', both assets and liabilities have to be expressed in terms of the monetary unit. In the case of physical assets, this means multiplying the quantity of the assets by some valuation coefficient which will turn it into a number of monetary units. Where the asset is constantly being bought and sold, the price, or ratio of exchange, is generally used as a valuation coefficient. In the case of fixed capital, the value is usually reckoned by taking an original purchase price and depreciating it over time by various methods, either at a

constant percentage rate or at a constant amount per year. This figure is very arbitrary in any case and in periods of inflation and deflation becomes extremely misleading. Inflation tends to increase accounting profits because fixed capital tends to be undervalued.

Another element in the situation is that all profit-making involves buying something at a certain price or cost at one time and selling it at a later time. If in the time interval all prices have risen, there is a spurious profit, which is not really represented by purchasing power. Thus there is much to be said for having a profit figure indexed, although the technical difficulties in this have so far prevented very much application of this principle. Inflation, therefore, produces illusory high profits; deflation, likewise, produces illusory low profits. This happened in the Great Depression, when accounting profits in 1932 and 1933 were negative. Unfortunately, it is accounting profits rather than real profits which tend to govern business expectations and decisions.

Beyond accounting, assets and liabilities make a very important contribution to the understanding of both the description and the dynamics of the economic system. Every liability is or should be an asset in some other balance sheet, for every debt is an asset to the creditor and a liability to the debtor. When we sum all the balance sheets in society, therefore, we should come out with an overall balance sheet that consists merely of real assets on one side and the total net worth of the society on the other. There is some question as to whether we should include money of various kinds in real assets. Bank deposits, of course, are assets to the holder and liabilities to the bank, so if we sum all assets, including banks, deposits would disappear. Even paper money is in a certain sense a liability of the government, although it is not usually reckoned as such, for it has to be accepted by government in payment of taxes. An important proposition follows from the concept of the aggregate balance sheet, that an increase in net assets, that is, investment, will produce an increase in the total of net worth, which is profit. This may be offset by other events. This is an important clue, however, to the dynamics of a great depression, which exhibits positive feedback: a decline in investment produces a decline in profits, a decline in profits produces a further decline in investment, a further decline in profits, and so on. This is clearly what happened between 1929 and 1933 in the capitalist world.

The relation of assets and liabilities to income, production and consumption is very important. Real assets can be regarded as a kind of ecosystem of goods, with the stock of each good representing a population. Production is then equivalent to births, consumption to deaths. Production minus consumption is the increase in the total stock of a particular good. The net national product is equal to the total production of goods, which is equal to the total consumption, plus an increase in the total stock of goods, just as an increase in any population is equal to the number of births minus the number of deaths in a given period.

Production is a function of the size and structure of real assets themselves, which is particularly clear if we include the value of the human bodies and minds (i. e. human capital) in the total, as ideally we should. Economists have an unfortunate way of regarding households as a kind of black box outside the economy proper. Actually they are very much a part of it, and household capital – houses, furniture, automobiles, clothing, and so on – is very close to half of the total in a modern society. When we fly over a city we see far more houses than factories. If we compare the capital around us at our workplace with the capital around us in our home, for a considerable part of the population the home capital is much larger than the capital at work.

Another very important problem is the contribution of assets, particularly household assets to economic welfare. There is a long tradition in economics that regards consumption as the main method of measurement of riches. It is clear, however, that we get most of our satisfaction from the use and enjoyment of assets rather than from their consumption. *I* get no satisfaction out of the fact that my car, house and clothing are wearing out. What *I* get satisfaction out of is using them. An increase in durability, especially of household capital, therefore, is an addition to economic welfare. This is a point much neglected by economists. Consumption, then, can usually be seen as a

bad thing, and production as what is necessary to offset it. There are exceptions to this rule. We like eating. We like the activity of producing in itself, even though it involves the using up of raw materials and so on. Thus the economic welfare function would include both assets of all kinds and certain forms of production and consumption, that is, income. Economists have often confused consumption with household expenditure or purchases, again because they regard the household as outside the economy. In modern society this can be very misleading, for household purchases are governed in no small degree by the depreciation of household capital to the point where it has to be replaced, so this depreciation is a very important aspect of consumption and income. Household purchases are exchange, not consumption. The production of assets include households also tends to be neglected, and it is an important part of the total economy in terms of cooking, mending, painting and repairing. The household has a balance sheet of assets and liabilities just as much as a business does and cannot be understood without it.

Human capital, both in terms of assets and liabilities, is a concept which has achieved some recognition. Economic development is primarily a process in human learning and the increase in human capital. A natural catastrophe or a war which destroys physical capital is restored remarkably quickly if the human capital remains intact and the knowledge and the know-how are unimpaired. We often do not realize that an enormous destruction of capital takes place every year just by depreciation and consumption. Even spectacular disasters are often just a relatively small addition to this annual destruction. The fact that some human beings have a negative human capital, both for themselves and for society, cannot be overlooked, though our social accounting system is ill-equipped to deal with this problem. In political decisions, however, we do recognize it. The criminal justice system is at least intended to diminish negative human capital; the educational system, to increase positive human capital. The fact that there is very little capital accounting in government means that considerable parts of its activity, like unilateral national defence organizations, do not really have a 'bottom line', and their value is usually assessed in non-economic terms, which can easily lead into catastrophic mistakes of judgement.

See Also

- ▶ Accounting and Economics
- ▶ Double-Entry Bookkeeping

Assignment Problems

Martin Beckmann

Suppose each member i from one class of objects (persons, firms) $i = 1, \ldots, n$ is matched with one object j from another class of equal size (jobs, locations) $j = 1, \ldots, n$ and the economic outcome is measurable in money terms a_{ij}. Let $x_{ij} = 1$ when object i is assigned to object j and $x_{ij} = 0$ otherwise. The payoff of this matching is then $\sum_{ij} a_{ij} x_{ij}$. It represents gross profits (profits before wages or rents) in the assignment of persons to jobs and of firms to locations.

In the personnel or plant assignment problem this is to be maximized subject to the constraints that x_{ij} be integer and that

$$\sum_{j=1}^{n} x_{ij} = 1 \tag{1}$$

and

$$\sum_{i=1}^{n} x_{ij} = 1 \tag{2}$$

This *Linear Assignment Problem* (Thorndike 1950; Von Neumann 1953; Koopmans and Beckmann 1957) represents the simplest type of an allocation problem involving indivisible resources.

Since one of the constraints is redundant and a feasible linear programme can always be solved with no more positive variables than active constraints, an argument by induction shows that the

integer constraints can be dropped so that the assignment problem becomes a special case of the transportation problem in linear programming (one unit to be removed from every point i and to be received at every point j). Even when partial assignments are meaningful (as in the case of assigning persons to jobs) an optimal assignment always exists that is a matching of persons with full-time jobs. This remains true when the constrains are relaxed ($\leqq$ instead of $=$) and when the number m of objects i is unequal to the number n of objects j.

An important implication of the fact that the linear assignment problem is a linear programme is the existence of efficiency prices p_i, q_j that characterize and sustain the solution.

$$p_i + q_j \leqq a_{ij}, \text{ and } `=' \text{ when } x_{ij} > 0. \quad (3)$$

In the personnel assignment problem this means the following: the gross profits a_{ij} of an optimal assignment are split into wage p_i and job rent q_j, and this is the highest return that either labour or job owner can earn,

$$\begin{aligned} p_i &= \max\left[a_{ij} - q_j\right] \\ q_j &= \max\left[a_{ij} - p_i\right] \end{aligned} \quad (4)$$

In the locational assignment problem p_i is the firm's net profit or rent and q_j the location rent. An optimal assignment is thus sustained by competitive markets in which p_i and q_j are charged as competitive prices, for any non-optimal assignment would not earn these wages and rents and thus incur a loss. When the number of objects to be matched is equal, the efficiency prices p_i,q_j, contain an arbitrary constant that may be added to all p_i and subtracted from all q_j. When all gross profits a_{ij} are positive, then a system of positive prices p_i, q_j exists. When there are more locations than firms, however, the efficiency prices of the non-occupied locations are zero and the arbitrariness disappears.

The dual problem requires one to find a minimal sum of wages and rents that covers all possible assignments

$$\min_{p_i, q_j} \sum_{i=1}^{n} p_i + q_i$$

such that $p_i + q_i \geqq a_{ij}$ for all $i, j = 1, \ldots, n$.

Suppose we arrange persons in the order of decreasing wages p_i. Then the optimal assignment results when we let persons choose jobs among the remaining vacancies in this order. When all are allowed to bid, however, the payoffs of the more attractive jobs must then be handicapped by job rents until an equilibrium is found in which every job attracts one and only one interested bidder. This is the person for whom this job realizes his comparative advantage (the absolute advantage as measured by the payoff a_{ij} is achieved only by that person who secures the highest wage). A person who scores higher on every job than another person will receive a higher competitive wage than the other person.

These results for the linear assignment problem apply also when multiple copies of the same job (machine) or of the same (type of) person are present.

An interesting variant is the room-mate problem where the objects come from the same set and the set contains an even number. There is then a single constraint

$$\sum_{j=1}^{n} x_{ij} + x_{ji} \leq 2 \quad (5)$$

which has a feasible solution where n is even.

Notice that each assignment is counted twice. The efficiency condition is

$$p_i + q_j \leqq a_{ij}, \text{ and } `=' \text{ when } x_{ij} > 0. \quad (6)$$

Any fractional assignment would now generate a closed chain of positive x_{ij} which can always be broken by decreasing some x_{ij} while increasing some other x_{jk} until no chains are left, resulting in an integer assignment once more.

The dual problem $\min_{\mu i} \sum_{i=1}^{n} \mu_i$ such that $p_i + p_j \geqq a_{ij}$ all $i, j = 1, \ldots, n$ is clearly feasible, so that when (5) is also feasible an optimum solution exists.

The linear assignment problem ignores interdependencies among the pairs formed by the assignment. Such interdependencies exist, however, in the location example when 'linkages' occur between the different plants through the exchange

of intermediate commodities. Let $b_{kl} \geqq 0$ denote the commodity flow in weight units from plant k to plant l and assume these numbers to be technical constants independent of location. Let c_{ij} denote the distance from location i to location j, measured in transportation costs incurred in moving one weight unit from i to j, and assume transportation costs for a commodity flow to be proportional to weight times distance. When plants k and l are assigned to locations i and j respectively, the total transportation cost incurred is

$$\sum_{i,j,k,l} x_{ik} b_{kl} x_{ij} c_{ij} \tag{7}$$

An optimal assignment is now one that minimizes (7) subject to the constraints (1) and (2). The minimand is quadratic and not concave. Relaxing the integer constrains will always result in fractional solutions, and it is the fractional solutions that would be sustained by the efficiency prices that are market prices. It follows that in general

> no price system on plants, on locations, and on commodities, in all locations that is regarded as given by plant owners and landlords will sustain any assignment. There will always be an incentive for someone to seek a location other than the one he holds. In the case of plants on the drawing board, competitive choices cannot be induced or sustained by such a price system. In the case of actual establishments already located the cost of moving is the only element of stability in the technological circumstances we have assumed. Without such a break on movement there would be a continual game of musical chairs. Whatever the assignment, prices of intermediate commodities and rents on locations cannot be so proportioned as to give no plant an incentive to seek a location other than the one it holds. (Koopmans and Beckmann 1957, p. 70)

Examples illustrating this have in fact been constructed. This is a disturbing case of market failure in the face of 'externalities', the externalities of the transportation costs incurred by others that result from the locational choice of any particular plant owner.

Mathematically the quadratic assignment problem turns out to be of the nonpolynomial type where the number of computations is not bounded by any polynomial function of the size of the problem (the number of plants), but workable algorithms have been developed (Graves and Whinston 1970; Geoffrion and Graves 1976; Reiter and Sherman 1962).

Suppose now that there is no cardinal measure for the outcome of a matching but only a preference ordering on the sets of agents, for example, medical students and hospitals (Roth 1984), or men and women (Gale and Shapley 1962). When each agent of one set is assigned to at most one agent of the other set, this is known as the marriage problem; if to more than one as the college admissions problem.

The marriage problem is defined by two disjoint sets of men $M = \{m_1,\ldots, m_n\}$ and women $W = \{w_1,\ldots, w_n\}$; each man has a strict preference ordering over the set $W \cup \{u\}$ of women, where u represents the possibility of remaining unmarried; and each woman has a strict preference ordering over the set $M \cup \{u\}$.

Thus each agent can compare the desirability of marrying a potential assignment from the opposite sex or of staying unmarried.

Let $w_j P(m) w_k$ denote the man m prefers woman j to woman k and $m_j P(w) m_k$ that woman w prefers man j to man k. An outcome of the marriage problem is an assignment $w = x(m)$ of women to men and of men to women $m = y(w)$ such that $w = x(m)$ if and only if $m = y(w)$. An outcome thus matches a subset of the women with a subset of the men in monogamous marriage and leaves the rest of the men and the women unmarried. It is called *individually rational* if no woman prefers being unmarried to the assignment $y(w)$ and no man prefers being unmarried to the assignment $x(m)$. An outcome is called *unstable* if it is not individually rational or if there exists a woman w and a man m who prefer each other to their assignments $y(w)$ and $x(m)$, $mP(w)$ $y(w)$ and $wP(m)$ $x(m)$. An assignment that is not unstable is called *stable*.

The set of stable assignments is the core $C(P)$ of the game with the following rules: any woman and any man marry if and only if they both agree and may remain unmarried if they prefer. This core is not empty: there always exists a stable outcome. For instance, let men be arranged in a fixed but arbitrary sequence and let the first man propose in the order of his preference until a woman accepts

him or he has exhausted the set W. The next man proposes in the order of his preferences to the women who have not yet married and so on. A man who precedes another man in this sequence would never prefer the other man's wife, or else the other man's wife would not prefer him.

For any man m the set of achievable assignments is the set $A_m(P) = \{x(m) | x \text{ is in } C(P)\}$ which represents the set of women to whom marriage is achievable with a stable outcome or if empty, the unmarried state. The set $A_w(P)$ of achievable assignments for women is defined analogously.

Proposition

The set $C(P)$ of stable outcomes of the marriage problem contains a M-optimal stable outcome x^* with the property that, for every man m in M, $x^*(m)$ is man m's most preferred achievable assignment; that is $x^*(m)$ $R(m)$ $x(m)$ for any other stable outcome x. Similarly, it contains a W-optimal stable outcome y^* such that $y^*(w)$ $R(w)$ $x(w)$ for every woman w and any stable outcome x. Thus all men are in agreement that x^* is the best stable outcome. By symmetry all women agree that y^* is the best stable outcome.

The M-optimal stable outcome has the following property reminiscent of Pareto-optimality: there is no outcome preferred by all men to the M-optimal stable outcome x^*. There is no outcome preferred by all women to the W-optimal stable outcome y^*. Thus even among unstable outcomes none is preferred by all men to the M-optimal stable outcome (similarly for women). An algorithm to discover x^* was proposed by Gale and Shapley (1962). In the first round each man proposes to the woman he ranks first. A woman rejects all proposals but one and accepts tentatively the man she prefers most among those who have proposed to her. In the second round the rejected suitors propose to their second choice. A woman may now jilt her first acceptance if she receives a proposal she prefers. This process continues until all men have been accepted or have exhausted their choice. This process represents a fair approximation to current practice in the US (i.e. 'sequential monogamy').

What interest does an agent have to reveal his true preferences which are only known to himself? Suppose some known procedure is applied by a planning board to produce a stable outcome with respect to the stated preferences of the agents, to be called a stable matching procedure. Any stable matching procedure gives rise to a game in which each agent's strategies are the preference orderings he/she might state.

There is no stable matching procedure which makes it a dominant strategy for all agents to state their true preferences. However,

> the matching procedure that yields the M-optimal stable outcome $x^*(P)$ for any stated preference P makes it a dominant strategy for every m in M to state his true preferences in the marriage problem. Similarly, a procedure that always yields $y^*(P)$ makes it a dominant strategy for every w in W to state her true preferences. (Roth 1985, p. 280)

Stable outcomes exist also for the college admissions problem, but their properties are somewhat weaker (Roth 1985). Also, every solution to the linear assignment problem is stable.

See Also

- ▶ Indivisibilities
- ▶ Integer Programming

Bibliography

Dubins, L.E., and D.A. Freedman. 1981. Machiavelli and the Gale–Shapley algorithm. *American Mathematical Monthly* 88: 485–494.

Gale, D., and L. Shapley. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69: 9–15.

Geoffrion, A.M., and G.W. Graves. 1976. Scheduling parallel production lines with changeover costs: Practical application of a quadratic assignment–LP approach. *Operations Research* 24(4): 595–610.

Graves, G.W., and A.B. Whinston. 1970. An algorithm for the quadratic assignment problem. *Management Science* 17: 453–471.

Koopmans, T.C., and M. Beckmann. 1957. Assignment problems and the location of economic activities. *Econometrica* 25(1): 53–76.

Reiter, S., and G.R. Sherman. 1962. Allocating indivisible resources affording external economies or diseconomies. *International Economic Review* 3: 108–135.

Roth, A.E. 1982. The economics of matching: Stability and incentives. *Mathematical Operations Research* 7: 617–628.
Roth, A.E. 1984. The evolution of the labor market for medical interns and residents: A case study in game theory. *Journal of Political Economy* 92: 991–1016.
Roth, A.E. 1985. The college admissions problem is not equivalent to the marriage problem. *Journal of Economic Theory* 36: 277–288.
Thorndike, R.L. 1950. The problem of classification of personnel. *Psychometrika* 15: 215–235.
Von Neumann, J. 1953. A certain zero-sum two-person game equivalent to the optimal assignment problem. In *Contributions to the theory of games*, vol. 2, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.

Assortative Matching

Li Hao

Abstract

This article reviews the simple economics of matching by characteristics. The goal is to understand sorting patterns in the marriage market and other matching markets by focusing on the nature of the gain from match and the mechanism of the market force of competition.

Keywords

Assortative matching; Becker, G.; Beckmann, M.; Complements and substitutes; Incomplete information; Koopmans, T.; Marriage markets; Matching frictions; Matching markets; Mixed matching; Price discrimination; Self-matching; Substitutes and complements; Transferable utility; Wage heterogeneity, sources of

JEL Classifications

C78

In a marriage market the competition for spouses leads to sorting of mates by characteristics such as wealth and education. 'Positive assortative matching' refers to a positive correlation in sorting between the values of the traits of husbands and wives (matching of likes); 'negative assortative matching' refers to a negative correlation (matching of unlikes). While it has been long recognized that sorting of husbands and wives by characteristics occurs in all cultures and societies, economists have tried to understand sorting patterns in the marriage market and other matching markets by focusing on the nature of the gain from match and the mechanism of the market force of competition.

The Basic Framework

A simple framework to illustrate the economic approach to sorting in matching markets is a two-sided marriage market with an equal number of men and women, who differ in one-dimensional characteristics called 'type' and have common preferences for higher types over lower types. In positive assortative matching, the highest-type man mates the highest-type woman, and the second-highest-type man mates the second-highest-type woman, and so on. Negative assortative matching is between the highest-type man and the lowest-type woman, between the second-highest-type man and the-second-lowest type woman, and so on. We assume transferable utility and zero reservation utility from remaining single for each market participant. Then, the gain from a match can be represented by an increasing, positive-valued function f, which gives the match output $f(x, y)$ of any pair of type x man and type y woman. Consider two men, with types $x_H > x_L$, and two women, with types $y_H > y_L$. If type x_H and type x_L command the same price in terms of the utility transfer they demand from the wife for the match, then both type y_H and type y_L would prefer the higher-type man because f is increasing in male type.

Competition for type x_H naturally leads to a higher price for type x_H than for type x_L. Whether the higher female type y_H can outbid type y_L for type x_H or vice versa depends on whether the male type and the female type are complements or substitutes in the match output function f. If

$$f(x_H, y_H) - f(x_H, y_L) > f(x_L, y_H) - f(x_L, y_L), \quad (1)$$

then the male type and the female type are complementary, because the marginal product of the female type is greater when matched with a higher male type (the left-hand side of inequality (1) than with a lower male type (the right-hand side of (1)). In this case, type y_L is willing to offer type x_H at most $f(x_H, y_L)$-$f(x_L, y_L)$ more than she offers type x_L, but by inequality (1) this difference is smaller than $f(x_H, y_H)$-$f(x_L, y_H)$, which is the most type y_H is willing to offer. Thus, type y_L will be outbid by type y_H for type x_H when the male type and the female type are complements. Since the argument is valid for any two pairs of men and women, the competition for spouses must lead to positive assortative matching. Conversely, if inequality (1) is reversed, male type and female type are substitutes. A lower female type can outbid a higher type for any male type, and the competition for spouses leads to negative assortative matching.

The differentiable version of inequality (1) is

$$\frac{\partial^2 f(x,y)}{\partial x \partial y} > 0. \tag{2}$$

Conditions (1) and (2) are commonly referred to as the (strict) 'supermodularity' condition of the match output function f. See Topkis (1998) for a comprehensive mathematical treatment of supermodularity, and Milgrom and Roberts (1990) and Vives (1990) for applications in game theory and economics.

Inequality (1) can be rewritten as

$$f(x_H, y_H) + f(x_L, y_L) > f(x_L, y_H) + f(x_H, y_L). \tag{3}$$

Condition (3) suggests that positive assortative matching maximizes the sum of match outputs in the marriage market when male type and female type are complements in the match output function. This result is a direct application of Koopmans and Beckmann's (1957) theorem of equivalence between efficient matching, which maximizes the sum of match outputs among all feasible pairwise matchings, and competitive equilibrium matching, which obtains when each woman y takes as given a schedule of utility transfers $u(x)$ to men and chooses the male type that maximizes her utility. Competitive equilibrium matching can also be obtained as each man x takes as given a schedule of utility transfers $v(y)$ to women and chooses the female type that maximizes his utility. Shapley and Shubik (1972) model the marriage market with transferable utilities as a cooperative game. They show that a pair of transfer schedules that support an equilibrium matching correspond to the core of the game, so that no pair of a man and a woman not matched in equilibrium can form a blocking coalition that produces a match output greater than the sum of their respective transfers.

Applications of Assortative Matching

The results of Koopmans and Beckmann (1957) and Shapley and Shubik (1972) are obtained in a matching market without any hierarchical ordering of types. By introducing one-dimensional, heterogeneous types, Becker (1973) seeks to explain why sorting of mates by wealth, education and other characteristics is similar in the marriage market. He constructs a household production function and derives condition (1) for each of the characteristics separately by considering how the characteristic affects household output while holding other characteristics fixed. Becker's model can accommodate dissimilar sorting of mates by some characteristics as well; for example, negative assortative matching by wage rates may arise because the benefits from the division of labour within a household can make the earning abilities of the man and the woman substitutes for each other.

Sattinger (1980) uses condition (2) to explain why the distribution of earnings of workers is skewed to the right relative to the distribution of their measured skills. In a market that matches a continuum of workers with different skills to a continuum of positions of different capital investment, the distribution of earnings would have the same shape as the distribution of skills if matching is random. In Sattinger's theory of differential rents, positive assortative matching of worker skill and job capital investment occurs because skill and capital investment are complements. In this case, the distribution of earnings will not resemble the distributions of outputs of workers at a job with the average capital investment. Instead,

workers with higher skills are paid more than those with lower skills both because they are more productive at any job and because they occupy positions with greater capital investments. Formally, in equilibrium the wage schedule u satisfies the first-order condition of type y's maximization problem of choosing x to maximize $f(x, y)$- $u(x)$

$$\frac{\partial f(x, m(x))}{\partial x} = u'(x),$$

where $m(x)$ is the capital investment of the job occupied by the worker with skill x in equilibrium. It can shown that condition (2) and positive assortative matching imply that $f(x, y)$- $u(x)$ is concave in x at $x = m^{-1}(y)$, so the second-order condition is satisfied for each y. The first-order condition implies that the worker's wage increases at the rate of the marginal product of the worker's skill x at his equilibrium job, so that the rate of increase of u is augmented by the complementarity (condition 2) and positive assortative matching ($m'(x) > 0$). Therefore, with positive assortative matching, the distribution of earnings will be positively skewed relative to the distribution of skills.

Kremer (1993) highlights the role of positive assortative matching in economic development. In his model of a one-sided, many-to-many matching market, each firm consists of a fixed number of workers, each employed for a production task. Workers have different skills, with a higher-skilled worker less likely to make mistakes in performing his task. Condition (1) is assumed to capture the complementarity among worker skills in the sense that the production process of a firm requires completion of each task without mistakes. Self-matching obtains in equilibrium where each firm employs workers of identical skills. Kremer uses this form of positive assortative matching to explain the large wage and productivity differences between developing and developed countries that cannot be accounted for by their differences in levels of physical or human capital.

Self-matching will generally be inefficient and will not occur in equilibrium if production tasks in a firm differ in skill requirements. In Kremer and Maskin (1996), a firm consists of two workers with a match output function $f(x, y)$ that satisfies the supermodularity conditions (1) and (2) but is asymmetric in that $f(x, y) > f(y, x)$ for any $x > y$. The interpretation of the asymmetry is that the first argument in f represents the skill of the worker who does the manager's job, while the second argument represents the skill of the worker who performs the assistant's job. In any given firm, it is optimal to make the higher-skilled worker the manager and the lower-skilled worker the assistant, but it is no longer generally true that self-matching maximizes the total match outputs. Indeed, we can have

$$2f(z_H, z_L) > f(z_H, z_H) + f(z_L, z_L) \qquad (4)$$

for some $z_H > z_L$, so that two firms each with the higher type z_H as the manager and the lower type z_L as the assistant produce more in total than two firms with the manager and the assistant having the same skill level. Note that inequality (4) does not contradict inequality (3) due to the asymmetry in f. Mixed matching may do better than self-matching because it can be more important to exploit the asymmetry in the match output function and have each high-skill worker as the manager of a firm than to exploit the complementarity in f and have one high-skill worker as the assistant to the other high-skill worker. Kremer and Maskin find that efficient matching in their model depends on the skill distribution in the matching market, because the trade-off between the asymmetry and the complementarity in the match output function depends on the relative scarcity of high-skilled workers.

Frictions in Matching Markets

Assortative matching may be hindered by the presence of frictions in the matching market. For example, if there is a moral hazard problem in producing the match output by each matched pair, transferability of utilities will be restricted by incentive compatibility constraints. Legros and Newman (2002) discuss this and other examples of transaction costs, and find that equilibrium matching in these examples can be inefficient. Frictions can also arise due to incomplete information about type. Roth and Xing (1994) provide

detailed descriptions of labour markets for entry-level professionals (such as lawyers and medical interns) in which early matches are sometimes made before complete information about matching characteristics, such as qualifications of job candidates and desirability of job positions becomes available. The complementarity in the match output function between the type of the applicant and the type of the job implies that there will be matching efficiency loss if matches are formed before the uncertainty about types is resolved. If all market participants are risk neutral, this efficiency loss is sufficient to rule out early matches as applicants compete for job positions. However, when some participants are risk averse, early matches provide them with some insurance against the payoff risks associated with late matches formed after complete information about types becomes available. Li and Suen (2000) apply competitive equilibrium analysis to the early matching market to determine the pattern of early matching, the terms of early matches, and the distribution of benefits in the early market. Early matching need not be positive-assortative in terms of expected type. Higher expected types of workers may face greater payoff risks from late matches due to the complementarity in the match output function. In this case, they may be willing to match with lower expected types of jobs to insure against the risks, while owners of higher expected types of jobs are content with waiting for late matches if they are risk neutral.

Private information about type may also result in frictions in the matching market. For example, many users of Internet dating agencies complain about the problems of misrepresentation and exaggeration by some users in the information they provide to the agencies. This problem arises because current matching services adopt a uniform pricing policy, and this in practice results in almost random matching. Damiano and Li (2007) point out that the complementarity in the match output function implies a version of the standard single-crossing condition in mechanism design problems, and an intermediary can use price discrimination to improve matching efficiency and generate greater revenue. They consider the problem of a monopoly matchmaker that uses a pair of fee schedules to sort different types of agents on the two sides into exclusive meeting places. The revenue-maximizing sorting need not be positive assortative (that is, efficient in the first-best sense). Conditions necessary and sufficient to recover positive assortative matching require that the complementarity in the match output function to be sufficiently strong to overcome the incentive cost to the matchmaker of eliciting private type information.

Matching frictions can arise also because finding type information about potential partners takes time or involves costly effort. In the search and matching framework, each market participant randomly meets a currently unmatched agent from the other side of the market, and decides whether to form a match or to search again in the next period. Search is costless, but agents must trade off the benefit from starting to produce with the encountered partner right away against the opportunity cost of waiting for a better partner. With an exogenous probability of separation of matched agents who then re-enter the market, Shimer and Smith (2000) characterize the stationary search and matching equilibrium where the matching decisions of each type and the type distributions of unmatched agents are time-invariant. Types x and y in an agreeable match are assumed to use the Nash bargaining solution to split the net surplus, defined as the match output $f(x, y)$ minus the sum of the (endogenous) continuation payoffs $g(x)$ to x and $h(y)$ to y as unmatched agents. Shimer and Smith modify the definition of positive assortative matching in the frictionless world to allow for set-valued mutually agreeable matches. The match set of a type x is the intersection of the set of types that type x agrees to match with and the set of types that agree to match with x. In Shimer and Smith's definition, matching is positive-assortative where, if for any male types $x_H > x_L$ and female types $y_H > y_L$ such that y_H is in the match set of x_L and y_L is in the match set of x_H, then y_H is in the match set of x_H and y_L is in the match set of x_L. When match sets are convex, positive assortative matching requires the lowest and the highest type of the match set to be increasing in x. However, match sets need not be convex

even though the match output function is supermodular. This is because the net surplus $f(x, y)$-$g(x)$- $h(y)$ is not necessarily quasi-concave in y for fixed x, so one cannot to say anything about how match sets vary across different x. Shimer and Smith provide conditions on f in addition to supermodularity to ensure convexity of match sets and re-establish positive assortative matching in a stationary equilibrium.

The stationary search and matching equilibrium does not capture the dynamics of matching in markets where there is no entry of a new cohort in each period and each matched pair receives their match output after the market closes for all participants. For example, many entry-level markets for professionals (such as academic economists) are organized around annual recruitment cycles. In these markets, matches are formed sequentially without centralized matching procedures. Damiano et al. (2005) consider such markets by constructing a two-sided, finite-horizon search and matching model with heterogeneous types and complementarity between types. The quality of the pool of potential partners deteriorates as agents who have found mutually agreeable matches exit the market. When search is costless and all agents participate in each matching round, the market performs a sorting function in that high types of agents have multiple chances to match with their peers. The matching efficiency measured by the total expected match outputs improves as the number of matching rounds increases; positive assortative matching is achieved if there are as many matching rounds as there are types. However, this sorting function is lost if agents incur an arbitrarily small cost in order to participate in each round. With a sufficiently rich type space relative to the number of matching rounds, the market unravels as almost all agents rush to participate in the first round, and match and exit with anyone they meet.

See Also

- ▶ Marriage Markets
- ▶ Matching and Market Design

Bibliography

Becker, G. 1973. A theory of marriage: Part I. *Journal of Political Economy* 81: 813–846.

Damiano, E., and H. Li. 2007. Price discrimination and efficient matching. *Economic Theory* 30: 243–263.

Damiano, E., H. Li, and W. Suen. 2005. Unravelling of dynamic sorting. *Review of Economic Studies* 72: 1057–1076.

Koopmans, T.C., and M. Beckmann. 1957. Assignment problems and the location of economic activities. *Econometrica* 25: 53–76.

Kremer, M. 1993. The o-ring theory of economic development. *Quarterly Journal of Economics* 108: 551–575.

Kremer, M., and E. Maskin. 1996. Wage inequality and segregation by skill. Working Paper No. 5718. Cambridge, MA: NBER.

Legros, P., and A. Newman. 2002. Monotone matching in perfect and imperfect worlds. *Review of Economic Studies* 69: 925–942.

Li, H., and W. Suen. 2000. Risk-sharing, sorting, and early contracting. *Journal of Political Economy* 108: 1058–1091.

Milgrom, P., and J. Roberts. 1990. Rationalizability, learning and equilibrium in games with strategic complementarities. *Econometrica* 45: 101–114.

Roth, A., and X. Xing. 1994. Jumping the gun: Imperfections and institutions related to the timing of market transactions. *American Economic Review* 84: 992–1044.

Sattinger, M. 1980. *Capital and the distribution of labor earnings*. Amsterdam: North-Holland.

Shapley, L., and M. Shubik. 1972. The assignment game I: The core. *International Journal of Games Theory* 1: 111–130.

Shimer, R., and L. Smith. 2000. Assortative matching and search. *Econometrica* 68: 343–369.

Topkis, D. 1998. *Supermodularity and complementarity*. Princeton: Princeton University Press.

Vives, X. 1990. Nash equilibrium with strategic complementarities. *Journal of Mathematical Economics* 19: 305–321.

Assumptions Controversy

Lawrence A. Boland

Keywords

Assumptions controversy; Cairnes, J. E.; Induction; Instrumentalism; Methodology of economics; Mill, J. S.; New Deal; Operationalism; Perfect competition; Robbins, L. C.; Senior, N. W.; Testing

JEL Classifications
B4

Today, any reference to an 'assumptions controversy' immediately calls to mind the many critical reactions to Milton Friedman's famous 1953 essay. But historians of economic thought will also point out that there was an assumptions controversy going back to the mid-19th century involving John Stuart Mill, John Elliot Cairnes and Nassau Senior (for an excellent review of this 'old' assumptions controversy, see Hirsch 1980). This old controversy was mainly between Mill and Senior and was about whether economics was an empirical science or a hypothetical one. The controversy was mediated by Cairnes and ultimately decided in his favour. For Cairnes, economic theory was true 'because it rested on premises which were undeniably true' (Hirsch 1980, p. 105). But any application of theory can be compromised by 'disturbing causes' and so the application needed 'to be compared with the facts' to see just what disturbing causes needed 'to be added in specific instances to make theory and facts correspond' (1980, p. 105). According to Abraham Hirsch, Cairnes's position reigned for over three-quarters of a century.

Friedman's essay was defending the use of perfect competition assumptions in applied economics against criticism of the assumption of universal maximization. The critics could easily find support in the philosophy of science of the day that claimed science is concerned with propositions that are meaningful because they are verifiable. But Friedman argued that, even in science, assumptions did not have to be true – only the logically derived results matter and theory should be judged according to whether these work or are useful. Friedman even argued it was acceptable to use simple assumptions that were obviously false on the grounds that one's theory might otherwise be so complex as to be useless.

Ideology as Method

Given the strong objections of most economists of this period to Friedman's views on markets, the suspicion must arise that ideology accounted for much of the interest in his methodology (Boland 2003). In particular, in the 1960s when Keynesian policies were thought by most mainstream economists to be obviously correct, Friedman's advocacy of a very limited role for the government was seen as a throwback to before the programmes of US President Franklin Roosevelt's New Deal that many other people thought helped overcome the Great Depression. But ideological arguments are not what academia is about. Instead, if one objects to Friedman's methodology, one must provide philosophical or scientific arguments against it to win the day. So, between 1957 and 1971 the controversy raged, not in the field of ideology but in the fields of semantics and methodology.

Ideology aside, it is difficult to understand why anyone would see Friedman's position to be very strong. After all, as I argued in Boland (1979), one can easily see Friedman's methodological position as nothing more than an up-to-date version of Instrumentalism (see instrumentalism and operationalism). And as such, if one were to ask Friedman or any Instrumentalists to defend their methodology – the methodology that claims the truth status of assumptions do not matter, only whether possibly false assumptions are useful – their only defence is to say that the Instrumentalist methodology itself works and hence is useful. There does not seem to be any other possible defence. But leading critics prior to 1979 seemed to think telling criticism could be provided. Unfortunately, none of their critiques was logically successful even though many opponents of Friedman's ideology wished to think so. To be effective, criticism of a doctrine must be in terms that a proponent of that doctrine would accept. Changing terms or imposing different objectives for the doctrine will not yield an effective or fair critique. All of the famous critiques published in the 1950s, 1960s and 1970s failed in this way.

Friedman's Instrumentalist Methodology

As explained in Boland (1979), any theory, in terms of Friedman's viewpoint, is an argument for some given propositions or towards specific

predictions. As such a theory consists only of a conjunction of assumption statements, that is, statements, each of which is *assumed* (or asserted) to be true; in order for the argument to be sufficient it must be a deductive argument. To be logically sufficient, an argument must satisfy the requirements of what logicians call *modus ponens*. To do so means that *whenever* all of the statements that make up the argument are true, all logically derived statements *must* be true. But quantificational logic also requires that, for a sufficient deductive argument in favour of some proposition, at least some of the assumptions must be in the form of universal general statements (in the form: 'all X have property Y'). With these two requirements in mind it should be evident that no purely inductive argument (one consisting *only* of particular statements such as observation reports) can be sufficient. The reason is simply that there is no purely inductive logic that satisfies *modus ponens*; that is, no inductive argument can *guarantee* that whenever all of the statements or assumptions that make up the argument are true that the conclusions will necessarily be true. Philosophers call this the problem of induction. It is a problem because without an inductive logic one cannot prove the truth status of any needed assumption in the form of a universal general statement (for example, '*all* firms are profit maximizers'). Friedman's 1953 essay attempts to overcome this key methodological problem.

Friedman's method simply dismisses the need to know that one's assumptions are true before deriving one's conclusions. The argument of his essay is that we are explaining given observation statements (for example, statements about the state of the economy) that are known already to be true. This means that the only requirement for any explanatory theory is that it does logically entail the truth of the observation statements – hence it forms a sufficient argument in favour of those observation statements. Moreover, there is no claim that the assumptions of the theory are necessarily true – only that, if they are true, the observed statements would be true. In other words, it is the *sufficiency* of the argument formed by any theory's assumption that matters, not the *necessity* of the theory's assumptions. In this sense, theories are tools or instruments for deriving known true statements. The test of an instrument can be only whether it works or is useful. This view of the *role* of theories is the essence of the doctrine of Instrumentalism. Proponents of Instrumentalism seem to think they have solved the problem of induction by ignoring the truth status of assumptions and thus they also imply that *modus ponens* will be of limited use. This is because Instrumentalist methodology does not begin with a search for the true assumptions but rather for true or useful (that is, successful) conclusions. Instrumentalist analysis of the sufficiency of a set of assumptions always begins by assuming the conclusion is true and then asks what set of assumptions will do the logical job of yielding that conclusion.

The Failed Critiques

Any valid or fair criticism of an Instrumentalist argument can only be about the argument's sufficiency. As a result, to refute an Instrumentalist argument one must show that the theory in question is insufficient, and thus inapplicable. The failure to recognize the logical requirements of any refutation of Friedman's 1953 methodology led to several failed critiques that nevertheless perpetuated the assumptions controversy. The first prominent shots fired in the assumptions controversy were by Tjalling Koopmans (1957) and Eugene Rotwein (1959), and the last – before the pot was stirred up again by Boland (1979) – was by Louis De Alessi (1971). In between were the critiques by Paul Samuelson (1963), Jack Melitz (1965) and Donald Bear and Daniel Orr (1967). As explained in Boland (1979), none of them dealt fairly or effectively with the Instrumentalism underlying Friedman's methodology as presented in his 1953 article. It should be acknowledged that the title of his article ('The methodology of positive economics') can be misleading. However, most misunderstandings are likely the result of his introduction, where he seems to be giving another contribution to the traditional discussions about methodology. Traditional discussions were

about issues such as the verifiability or refutability of truly scientific theories. But Friedman's essay does not do this. Instead, he actually gives an alternative to that type of discussion.

Following traditional discussion, Koopmans sees all theorists seeking to develop or analyse the 'postulational structure of economic theory' so as to obtain 'those implications that are verifiable or otherwise interesting' (1957, p. 133). Unlike Friedman's essay, which presumes that what one assumes depends on one's purposes, Koopmans presumes all theories are directly analysable independently of their uses. Koopmans's critique of Friedman's essay is based on a restatement of Lionel Robbins's methodological position (1935) which itself seems to be a restatement of what Cairnes argued. Koopmans's basic concern (but not Friedman's) is the sources of the basic premises or assumptions of economic theory. For the followers of Robbins, the assumptions of economic analysis are promulgated and used *because* they are (obviously) true. The truth of the assumptions is never in doubt. The only question is whether they are necessary for the mathematical derivation of the interesting implications.

Koopmans objects to Friedman's dismissal of the problem of clarifying the truth of the premises, the problem that Koopmans wishes to solve using mathematics. Koopmans is an inductivist and as such defines successful explanation as being logically based on inductively and observably true premises. Friedman does not consider assumptions or theories to be the embodiment of truth but only as instruments for the generation of useful (because successful) predictions.

In order to criticize Friedman's argument, Koopmans offers an *interpretation* of his own theory of the logical structure of Friedman's view. His interpretation contradicts Friedman's purpose (that *some*, but not necessarily all, conclusions need to be successful). It is most important to keep in mind that Friedman's methodology is concerned only with the *sufficiency* of a theory's set of assumptions. Koopmans falsely assumes that Friedman's methodology has a concern for *necessity*. In other words, Koopmans's theory of Friedman's methodology is itself void because (by Koopmans's own rules) at least one of its assumptions is false (for more, see Boland 1979, pp. 515–17).

Many self-proclaimed 'empiricists' accept the obviousness of the premises of economic theory. For them, the truth of one's conclusions (or predictions) rests *solely* (and firmly) on the demonstrable truth of the premises – and the presumption that one *must* also justify every claim for the truth of one's conclusions or predictions arrived at by *modus ponens*. Needless to say, such empiricists do not see a problem of induction. Friedman clearly does, and in this sense he is not an orthodox empiricist (despite the term 'positive' in his title, which usually means 'empirical'). According to the empiricist critic Rotwein, Friedman is criticizing views such as his by claiming that they represent 'a form of naive and misguided empiricism' (Rotwein 1959, p. 555). Actually, Rotwein sees the thrust of Friedman's essay as a family dispute among empiricists.

Obviously, there is 'good' and 'bad' naivety. Good naivety exposes the dishonesty or ignorance of others. But Friedman's essay does not join with the empiricist's pretence that there is an inductive logic, one that would serve as a foundation for Rotwein's verificationist empiricism. Rotwein twists the meaning of 'validity' into a matter of probabilities so that he can use something like *modus ponens* (1959, p. 558). But *modus ponens* will not work with statements whose truth status is a matter of probabilities (see Haavelmo 1944), and thus Friedman is correct in rejecting this approach to empiricism (for more, see Boland 1979, pp. 517–18).

A more sophisticated critique of Friedman's methodology is the one by Bear and Orr (1967). They criticize only certain aspects while accepting others. In particular, they dismiss Friedman's Instrumentalism while simultaneously recommending what they call his 'as if' principle. Their reason is that they too accept the view that the problem of induction is still unsolved but they see his principle as an adequate means of dealing with that problem. Their main complaint is that Friedman erred by 'confounding... abstractness and unrealism' (1967, p. 188, n. 3). Each part of Friedman's argument is, of course, designed only to be *sufficient*, but they ignore this and just claim

Friedman's arguments against the *necessity* of testing and against the *necessity* of 'realism' of assumptions are both wrong. They go further to claim, 'all commentators except Friedman seem to agree that the testing of the whole theory (and not just the predictions of theory) is a constructive activity' (1967, p. 194, n. 15). However, this criticism is unfair because Friedman's concept of testing (as verifying) does not correspond to theirs. Of course, it is not always clear what various writers mean by 'testing', mostly because its meaning is too often taken for granted. Where Friedman sees testing only in terms of verification or 'confirmation', Bear and Orr appear to adopt Karl Popper's view that a *successful* test is a refutation (Bear and Orr 1967, pp. 189 ff.). In a similar vein, another critic, Melitz (1965, pp. 48 ff.), seems to be saying that a successful test is confirmation or disconfirmation. In both critiques, the logic of the criticism is an allegation of an inconsistency between *the critic's* concepts of testing and Friedman's rejection of the necessity of testing assumptions. The logic of such criticism may be valid, but in each case the criticism is based on a rejection of Instrumentalism even though it is an absolutely essential part of Friedman's essay. Consequently, the critics are wrong as the alleged inconsistency does not exist *within* Friedman's Instrumentalist methodology. Moreover, it is unfair for critics to assert criticisms only on the basis of an inconsistency between *their* concept of testing and Friedman's methodological judgements which are based on *his* concept (for more, see Boland 1979, pp. 520–1).

De Alessi (1965, 1971) offered more friendly criticisms. First, he meekly criticizes Friedman for seeing only *two* attributes of theories; a theory can be viewed as a language and as a set of substantive hypotheses. De Alessi says, 'Unfortunately, Friedman's analysis has proved to be amenable to quite contradictory interpretations' (1965, p. 477). And, like Koopmans's criticism, it is presumed that Friedman is relying on *modus ponens*. But Instrumentalism, by not requiring true assumptions, cannot use *modus ponens*. So, such a presumption is false.

In his later article, De Alessi says Friedman argues that some assumptions and conclusions are 'interchangeable'. De Alessi notes that such 'reversibility' of an argument allows it to be tautological. Moreover, whenever an argument is tautological, it cannot also be empirical, that is, positive. The logic of De Alessi's argument may be correct – but it is not clear that Friedman was indicating 'reversibility' of (entire) arguments with the term 'interchangeable'. The only methodological point Friedman was making was that the status of a statement's being an 'assumption' is not necessarily automatic.

The most celebrated criticism of Friedman's methodology was presented by Samuelson (1963) in his discussion of Ernest Nagel (1963). Samuelson claims that Friedman is in effect saying that a 'theory is vindicable if (some of) its consequences are empirically valid to a useful degree of approximation; the (empirical) unrealism of the theory "itself", or of its "assumptions", is quite irrelevant to its validity and worth' (1963, p. 232). Samuelson labels this the 'F-Twist'. And about this he says it is 'fundamentally wrong in thinking that unrealism in the sense of factual inaccuracy even to a tolerable degree of approximation is anything but a demerit for a theory or hypothesis (or set of hypotheses)' (1963, p. 233). But Samuelson admits that his characterization of Friedman's view may be 'inaccurate' – supposedly why he labelled it the 'F-Twist' rather than the 'Friedman-Twist'. Nevertheless, Samuelson willingly applies his potentially false assumption in his explanation of Friedman's view. His justification for using a false assumption is Friedman's own 'as if' principle. In this way, Samuelson argues that followers of Friedman's methodology must concede defeat if one can discredit or refute Friedman's view by using Friedman's view. Samuelson admits there is 'cheap humor' in this line of argument. Nevertheless, he is attempting to criticize Friedman by using Friedman's own methodology. But by Samuelson's own mode of argument, his assumption that attributes the F-Twist to Friedman is false and the attempt to apply this by means of *modus ponens* is thus logically invalid.

Surely it is illogical (and at best pointless) to criticize someone's view with an argument that gives different meanings to the essential terms. But this is just what the prominent critics

do. Similarly, using assumptions that are allowed to be false while relying on *modus ponens*, as Samuelson does, is also illogical. Beyond preaching to the choir, an effective criticism must deal properly with Friedman's Instrumentalism. Any criticism that ignores his Instrumentalism will be an irrelevant critique. For this reason, the critiques of Koopmans, Rotwein and De Alessi are clear failures. None of the famous critics was willing to straightforwardly criticize Instrumentalism.

Towards Resolving the Assumptions Controversy

The obvious critique that might succeed is to dispute the success of the observations that Friedman and his followers choose to explain by using his Instrumentalist methodology. For example, it is all too easy to find special cases where maximum dependence on the market can solve social problems. Of course, many people would still not accept Friedman's advocacy of policies involving minimum government if based only on selected examples. But any dispute about Friedman's policy views would open the door to straightforward ideological arguments on the floor of academia. Without this (or at least a critique of the positive claims that are claimed to underlie Friedman's policy views), the controversy will never be decided in favour of Friedman's critics other than to simply recognize – as argued in Boland (1979) – that the only justification for Instrumentalist methodology is a self-serving appeal to Instrumentalism itself. Surely this would be a weak if not dishonest defence.

See Also

► Instrumentalism and Operationalism

Bibliography

Bear, D.V.T., and D. Orr. 1967. Logic and expediency in economic theorizing. *Journal of Political Economy* 75: 188–196.

Boland, L. 1979. A critique of Friedman's critics. *Journal of Economic Literature* 17: 503–522.

Boland, L. 2003. Methodological criticism vs. ideology and hypocrisy. *Journal of Economic Methodology* 10: 521–526.

De Alessi, L. 1965. Economic theory as a language. *Quarterly Journal of Economics* 79: 472–477.

De Alessi, L. 1971. Reversals of assumptions and implications. *Journal of Political Economy* 79: 867–877.

Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*. Chicago: University of Chicago Press.

Haavelmo, T. 1944. Probability approach to econometrics. *Econometrica* 12(Suppl): 1–118.

Hirsch, A. 1980. The 'assumptions' controversy in historical perspective. *Journal of Economic Issues* 14: 99–118.

Koopmans, T. 1957. *Three essays on the state of economic science*. New York: McGraw-Hill.

Melitz, J. 1965. Friedman and Machlup on the significance of testing economic assumptions. *Journal of Political Economy* 73: 37–60.

Nagel, E. 1963. Assumptions in economic theory. *American Economics Review* 53: 211–219.

Robbins, L. 1935. *An essay on the nature and significance of economic science*. London: Macmillan.

Rotwein, E. 1959. On the methodology of positive economics. *Quarterly Journal of Economics* 73: 554–575.

Samuelson, P.A. 1963. Problems of methodology: Discussion. *American Economic Review, Papers and Proceedings* 53: 231–236.

Asymmetric Information

A. Postlewaite

The Arrow-Debreu model is the basic model in which the two classical welfare theorems of economics are expressed. Under quite general assumptions, it can be shown that, first, a competitive equilibrium allocation, or Walrasian allocation, is Pareto efficient (Pareto optimal); second, under somewhat different assumptions, any Pareto efficient allocation will be a competitive equilibrium allocation after some suitable redistribution of initial endowments. Implicitly or explicitly, the statement of the first welfare theorem assumes that all economic agents have the same information about all economic variables. This is not to say that uncertainty is ruled out; there may

be uncertainty as long as all agents are identically uncertain. If this assumption of symmetric information is violated, the competitive outcome will no longer be guaranteed to be Pareto efficient. The introduction of asymmetric information into various economic problems has given us new insight into how market failures might arise and whether there may be governmental, or other non-market, corrections which can improve welfare. Several examples illustrating this are given below.

There may be a good which can vary in quality and whose quality will be known only by the owner. As an example, one can think of the objects being sold as used cars. Potential buyers will realize that there are good and bad quality cars and will rationally pay a price based on the average quality. This means that some cars will be underpriced (the highest quality cars), but which ones will be known only to the owner. Some underpriced cars may be so much underpriced that the owners will not be willing to sell them at the price based on the average quality. But the withdrawal of these quality cars causes the average quality of the cars in the market to decrease and consequently, potential buyers will rationally lower the price they are willing to pay for a car randomly drawn from those remaining. This in turn may lead to another round of withdrawal of some of the better remaining cars and a further lowering of the price buyers will pay. In the extreme, the equilibrium of this process may have no cars sold even in the case that all would have been sold had the quality of goods been symmetrically known, that is, when either everyone or no one could determine the quality. This problem is essentially that analysed by Akerlof (1970).

Asymmetric information has been introduced into a labour-management model to illustrate how it may distort the optimal labour contract. Assume that the demand function facing the firm is known to the firm but not to the workers. An optimal contract would generally be characterized by a constant labour force and a variable wage, lower wages being associated with lower levels of demand. This may not be feasible given the asymmetry, however. The firm would announce that the state of demand is low regardless of the truth since this lowers its wage bill without cost. The optimal contract with the asymmetry will typically involve a lower amount of labour employed when the firm announces that demand is low. Since this is more costly when demand is high (and the marginal revenue product of an additional hour is high), than when demand is low, optimal contracts in the presence of this sort of asymmetric information often take this form. Rosen (1985) surveys the literature on this problem.

A third area in which asymmetric information has been successfully introduced into traditional economic problems is that of industrial organization. As an example, it can be assumed that there are several firms within an industry and that each may know more about its own cost structure than about its competitor's (or potential competitor's). Equilibria in such models conform better to what is generally believed to be involved in predatory pricing and limit pricing than equilibria in models without asymmetric information. (Examples of such arguments can be found in Milgrom and Roberts 1982a, b.) It has also been shown that if small amounts of asymmetric information are introduced into the finitely played prisoners' dilemma game and into the chain store paradox, the paradoxes associated with these games disappear (see, e.g., Kreps et al. 1982).

In public economics, models have been investigated in which individuals know their own valuation for public goods but know nothing about other individuals' valuations. These models provide explanations of how and why governments may want to provide public goods. These explanations improve upon the explanations provided by models without asymmetric information; in addition, they provide a clearer understanding of the nature of the improvement in welfare that a government can effect. (Bliss and Nalebuff (1984) gives an insight into the problem of public goods with asymmetric information.)

The above examples focus on positive models which encompass asymmetric information. That is, they provide models which depend upon asymmetries in information to explain phenomena which are generally believed, but which are difficult to reconcile with optimizing behaviour in the absence of such asymmetries. There is extensive use of asymmetric information in normative

models as well. We may want to devise governmental or non-governmental mechanisms to augment, alter, or replace markets; if so, we presumably want to do so in an 'optimal' manner, whatever notion of optimality we may want to rely on. To the extent that there is asymmetry among the agents in the economy in question, we must be able to predict the outcome after our augmentation, alteration or replacement in the face of this asymmetry. This approach has been used extensively in optimal taxation. Suppose one feels that a given amount of tax revenue must be raised and that it is fairest to raise more revenue from those who are most able (most productive). If the ability of an agent is known to himself but not to anyone else, this asymmetry of information has to be taken into account. We must maximize social welfare subject to the constraint that it must be in the individual's interest to reveal, indirectly or directly, these privately known abilities. In this manner, it is possible to derive characteristics of an optimal tax schedule under asymmetric information. In a similar manner we can determine the qualitative characteristics of other types of taxes to be levied in environments with asymmetric information. Atkinson and Stiglitz (1980) is an excellent reference to the literature in this area.

Similar normative models have been used to investigate the nature of optimal policy for many problems such as regulatory policy, anti-trust policy, monetary policy and other problems in which asymmetric information may play a role.

The common technique in analysing both normative and positive problems with asymmetric information is to model them as games with incomplete information and to use the Bayesian–Nash solution concept. This captures both the asymmetric information and the problems raised by economic agents sometimes having incentives to misrepresent the information they have. This modelling technique is not wholly satisfactory, however. Embedded in the technique is the assumption that the information structure is common knowledge. This is an assumption that while an agent may not know the exact information that another agent has, he knows the probability distribution of the information. Further, the second agent knows that the first knows this, the first knows that the second agent knows that he knows, and so on ad infinitum. The assumption that the information structure is common knowledge is extremely strong and the results of models using the assumption are correspondingly less convincing. Myerson (1979) is the standard reference here.

Much of the use of asymmetric information in economic models was motivated by a desire to understand seeming (Pareto) inefficiencies in particular market situations. The integration of asymmetric information into economic models accomplished this. In addition, the formalization of the asymmetry in the information among agents helped to clarify the notion of welfare in such circumstances as well. The question of whether or not a change from one allocation to another might make all agents better off is unambiguous in the case that there is no uncertainty. With uncertainty which is identical for all agents, it is also simple; each agent makes the comparison between the two allocations by taking the expected utility of the allocations using the commonly accepted probability distribution. When each agent has different information the problem becomes more complicated. Some agents may know that certain events cannot happen while others may not know this. What probabilities should be used to calculate an agent's expected utility – his own beliefs, those of the best informed agent, the totality of the information held by all agents or some entirely different probability? Holmstrom and Myerson (1983) provide a careful analysis of welfare judgements in the face of asymmetric information.

The introduction of asymmetric information into models in which agents behave strategically made it necessary to consider not only what agents knew, but what they thought other agents knew, what they thought other agents knew about what they knew and so forth. Addressing this directly resolved many of the dilemmas posed by welfare comparisons in an environment with asymmetric information.

See Also

► Adverse Selection
► Implicit Contracts

▶ Incentive Contracts
▶ Incomplete Contracts
▶ Moral Hazard
▶ Principal and Agent (i)

References

Akerlof, G. 1970. The market for lemons. *Quarterly Journal of Economics* 84: 488–500.
Atkinson, A., and J. Stiglitz. 1980. *Lectures on public economics*. New York: McGraw-Hill.
Bliss, C., and B. Nalebuff. 1984. Dragon-slaying and ballroom dancing: The private supply of a public good. *Journal of Public Economics* 25: 1–12.
Holmstrom, B., and R. Myerson. 1983. Efficient and durable decision rules with incomplete information. *Econometrica* 51: 1799–1820.
Kreps, D., P. Milgrom, J. Roberts, and R. Wilson. 1982. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory* 27: 245–252.
Milgrom, P., and J. Roberts. 1982a. Predation, reputation, and entry deterrence. *Journal of Economic Theory* 27: 280–312.
Milgrom, P., and J. Roberts. 1982b. Limit pricing and entry under incomplete information: An equilibrium analysis. *Econometrica* 50: 443–459.
Myerson, R. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47: 61–74.
Rosen, S. 1985. Implicit contracts: A survey. *Journal of Economic Literature* 23: 1144–1175.

Atomistic Competition

A. P. Kirman

This term was originally taken from the physical concept of matter as composed of atoms, the smallest irreducible elementary particles in a void. This idea, which originates with Democritus and Epicurus, was adopted in the 19th century by economists to convey two ideas. The first, which has persisted, is the notion that individuals are many and unimportant. This has led to an assimilation in the French literature of 'atomistic competition' to 'perfect competition'. However a second and more subtle idea was implied and received its clearest early expression in the work of Adam Smith. This is the concept of a society or economy as 'atomistic' rather than 'organic'. Thus it is the actions of many independent individuals which determine the evolution of the whole, rather than the collective organization of these individuals. This idea was contested in particular by Marx, whose position was the opposite of that of the utilitarians.

Paradoxically in modern terms the term atomistic competition is wholly inappropriate as a description of the perfectly competitive model. The mathematical idea which corresponds to perfect competition is that of an 'atomless measure space' of agents. This conveys accurately the idea that although no individual has any weight, collectively they can have positive weight or influence. An atom in this context is an individual who alone does have weight and thus can influence economic outcomes. This might correspond to a very wealthy agent, a firm or a monopolist (*see* "▶ Measure Theory"). Thus the appropriate modern term would be 'atomless competition'.

See Also

▶ Large Economies
▶ Measure Theory

Attwood, Thomas (1783–1856)

B. A. Corry

Keywords
Attwood, M.; Attwood, T.; Birmingham School; Chartism; Cobbett, W.; Full employment; Mill, J. S.; Paper currency; Ricardo, D.

JEL Classifications
B31

In British social and political history the name of Thomas Attwood is usually connected with the

Birmingham Political Union, of which he was a founder, and hence the part that movement played in the peaceful enactment of the great Reform Act of 1832. Later he was also associated with the Chartist movement. However, Attwood also has a place in the history of economic thought as an early exponent of anti-classical monetary and macroeconomic ideas and as the leading member of the so-called Birmingham School.

Thomas Attwood was born in 1783, the son of a banker and into whose profession he followed. From an early age he was also active in public affairs in the City of Birmingham. In 1811 he was elected high Bailiff of that town and the following year, with Richard Spooner (later to be another notable member of the Birmingham School) he represented Birmingham manufacturers' interests against the Orders in Council that had restricted UK trade with the USA and the Continent.

He was first drawn into monetary controversy by the depression that followed the ending of the Napoleonic wars in 1815. Birmingham was then an important manufacturing town and had become the centre of small arms manufacture during the wars. Hence the abrupt reduction in government demand had a quick and sharp effect on the local economy. Attwood was particularly incensed by the cavalier attitude adopted by some orthodox classical economists towards the distress brought about by the post-war depression. Ricardo, for example, expressed little knowledge of it and doubted the claims of Birmingham industrialists. Attwood's first pamphlet – *The Remedy* – appeared anonymously in 1816 and this was followed in 1817, under his own name, by *A Letter to Nicholas Vansittart on the Creation of Money, and its Action upon National Prosperity.*

Those early pamphlets give us the theme that was to dominate all of Thomas Attwood's writings in the field of monetary economics. His prime object was the abolition of the metallic standard and its replacement with a flexible, managed, currency which, he believed was essential for a full employment policy. Throughout his many subsequent writings he never wavered from this position.

In 1830 Attwood was a founder of the Birmingham Political Union for the Protection of Public Rights: its aim was to secure middle and lower class representation in the House of Commons and the Union played a crucial role in supporting the Grey administration during the passage of the Reform Bill of 1832. In the same year together with Joshua Scholefield he was returned unopposed as a Member of Parliament for the new Parliamentary Borough of Birmingham. He continued to agitate for further Parliamentary reform and in 1839 was a presenter of the mammoth Chartist Petition to Parliament.

His place in the Chartist movement was uneasy and ambiguous. He never endorsed the use of physical force that was advocated by some of the more extreme leaders of the movement. More fundamentally the central tenet of Attwood's monetary proposals – the introduction of an inconvertible paper currency – was utterly rejected by the Chartists who attacked what they termed 'rag botheration' (paper currency) as enthusiastically as Cobbett.

Attwood felt, and rightly so, that his monetary ideas were never taken seriously by the establishment and he undoubtedly suffered from what may be termed a persecution complex. He was for example, caricatured by Disraeli in the *Runnymede Letters* and by J.S. Mill in the *Currency Juggle.*

Attwood died in 1856 a disappointed man. Birmingham honoured him with a statue in Stephenson's Place (1859).

His brother Matthias also wrote some important pamphlets in monetary matters but never took up the extreme position of his brother Thomas.

Selected Works

1964. *Selected economic writings.* Edited with an Introduction by F.W. Fetter. London: LSE Reprints of Scarce Works on Political Economy.

Bibliography

Briggs, A. 1948. Thomas Attwood and the economic background of the Birmingham Political Union. *Cambridge Historical Journal* 9 (2): 190–216.

Checkland, S.G. 1948. The Birmingham economists 1815–1850. *Economic History Review*, Second Series 1(1): 1–19.
Corry, B.A. 1962. *Money, saving and investment in English economics*. London: Macmillan.
Wakefield, C.M. 1885. *Life of Thomas Attwood*. Printed privately.

Auctioneer

F. H. Hahn

Abstract

The auctioneer is a fictitious agent, introduced by Leon Walras, who matches supply and demand in a market with perfect competition. The process is called 'tâtonnement', finding the market clearing price for all commodities, resulting in general equilibrium. No actual trading occurs during this process. The concept of the auctioneer sidesteps the important question of the coordinating power of the price mechanism. There are in fact only a few special cases for which the auctioneer process leads the economy to an equilibrium.

Keywords

Auctioneer; Auctions; Competitive equilibrium; Keynesianism; Monopolistic competition; Perfect competition; Tâtonnement; Walras, L.; Walras's Law

JEL Classifications

D0

Walras (1874) introduced the idea of a tâtonnement to provide a theoretical account of the formation of equilibrium prices. This account was not meant to be taken descriptively but rather as a 'Gedanken Experiment'. It was hoped that its study would provide insights into the actual *modus operandi* of the price mechanism.

Consider an economy of H households, F firms and n goods. Let $p \in \Delta \subset R^n_+$, where p is a price vector and Δ the simplex. Given the endowments of households $\left(e^h \in R^n_+\right)$, $x^h - e^h$ is the net trade vector of household h where $x^h \in R^n_+$ is the vector of demand of household h. Assume that

$$x^h - e^h = \xi_h(p)$$

where $\xi h(p)$ is a continuous function from Δ to R^n. Let $y\,f \in R^n$ be an activity of firm f, where $y^f_i > 0$ is interpreted as 'the firm supplies good i' and $y^f_i < 0$ is interpreted as 'the firm demands good i as an input'. Let $y = \sum_f y^f$ and assume that

$$y = \eta(p)$$

is a continuous function from Δ to R^n. Then define

$$z = \sum \left(x^h - e^h\right) - y$$

which by our assumptions can be written as, say

$$z = \sum_h \xi h(p) - \eta(p) = \theta(p).$$

It is known that addition of budget constraints implies

$$p \cdot z = 0 \text{ all } p \in \Delta.$$

(Walras's Law). An equilibrium of the economy is $p^* \in \Delta$ such that

$$\theta(p^*) \leq 0.$$

It should be added that the net trades $\xi h(p)$ are assumed to be utility maximizing for each household under the budget constraint:

$$p \cdot \xi_h(p) \leq \sum_f \lambda_{hf}\left(p \cdot y^f\right)$$

where $1 \geq \lambda_{hf} \geq 0$, $\sum_h \lambda_{hf} = 1$, is the share of h in the profits of firm f. Similarly $\eta^f(p) = y^f$ satisfies for all f: $p \cdot \eta^f(p) \geq p \cdot yf$ all y which the firm can choose amongst.

A tâtonnement is now described as follows. A fictitious agent called the *auctioneer* announces $p \in \Delta$. Households now report to this auctioneer their desired net trades $[\xi_h(p)]$ and firms report to him their desired activities $[\eta^f(p)]$. From these reports the auctioneer can deduce $\theta(p)$. In its light he calculates a new price vector p' as follows:

$$\frac{p_i'}{\Sigma p_i'} = \frac{p_i}{\Sigma p_i} \text{ if } \theta_i(p) = 0 \text{ or if } \theta_i(p) < 0 \text{ and } p = 0 \frac{p_i'}{\Sigma p_i'} > \frac{p_i}{\Sigma p_i} \text{ if } \theta_i(p) > 0 \frac{p_i'}{\Sigma p_i'} < \frac{p_i}{\Sigma p_i} \text{ if } \theta_i(p) < 0.$$

He announces p' agents send back messages which allow him to calculate $\theta(p')$. The process continues until and if the rule for calculating a new price vector yields the preceding price vector. *No actual trading occurs* during this process.

The rule which we have supposed the auctioneer follows in changing his price announcement is only one of a number of possible ones. Indeed, it is not the one proposed by Walras. He supposed the auctioneer to concentrate on one market at a time; specifically he changes only one price. Suppose he changes the *i*th price. Then he changes it until, given all other prices which are held constant, the *i*th market is in equilibrium. (He assumed that there always is such a price and that it is unique.) Thereafter he moves on to the next market. Of course, this process may never terminate in an equilibrium.

In all of this one ought to specify what it is that the auctioneer knows. So far we have assumed that he does not know the function $\theta(p)$. If, however, he does know this function we may think of the auctioneer as being concerned to find a solution to $\theta(p) \leq 0$ for $p \in \Delta$. He is then no more than a programmer. In this case, for instance, he may adopt Newton's method (Arrow and Hahn 1971; Smale 1976). That is he proceeds as follows: Let $J(p)$ be the $(n-1) \times (n-1)$ Jacobian of the first $(n-1)$ excess demand functions $\left[\hat{\theta} = \theta_1(p)..., \theta_{n-1}(p)\right]$. The price of the nth good is set identically equal to unity (it is the numeraire). Then define $\hat{p} = (p_1, \ldots, p_{n-1})$ and let $\hat{q} = (q_1, \ldots, q_{n-1})$ solve:

$$\hat{\theta}(\hat{p}) - J(\hat{p})(\hat{q} - \hat{p}) = 0$$

where it is assumed that a solution exists:

$$(\hat{q} - \hat{p}) = J(\hat{p})^{-1}\hat{\theta}(\hat{p}).$$

The auctioneer now follows the rule: raise pi, if $q_i - p_i > 0$, lower pi if $q_i - p_i < 0$ if $q_i - p_i < 0$ and $p_i > 0$ and leave pi unchanged if either $qi = pi$ or $qi < pi$ and $pi = 0$. Under certain technical assumptions this way of calculating will lead the auctioneer to an equilibrium (see Arrow and Hahn 1971).

This example demonstrates that it is possible to think of a tâtonnement as a kind of computer program. If one adopts this view, however, one will certainly not be mimicking the invisible hand. For instance, in the Newton method the price change in any one market depends on the excess demand functions in all markets and that is not what any version of 'the law of supply and demand' stipulates. Moreover the proposal violates the supposed economy in information of decentralized economies – that is, much more is known to the auctioneer than can be known to any one agent. From the point of view of positive theory, therefore, this second interpretation of the auctioneer is not helpful, although it has found application in the theory of planning (e.g. Heal 1973).

Assuming that the auctioneer only knows aggregate excess demands at the announced p, it has been customary ever since a famous paper by Samuelson (1941, 1942) on Hicksian stability to formulate the rule followed by the auctioneer dynamically. For instance:

$$\frac{dp_i}{dt} = 0 \quad \text{if } \theta_i(p) < 0 \text{ and } p_i = 0$$
$$\frac{dp_i}{dt} = k_i\theta_i(p) \quad \text{otherwise with } k_i > 0.$$

Even if this process leads to p^* it will do so only as $t \to \infty$. This is awkward since no one is allowed to trade while the process is still in motion. Some economists have by-passed this by saying that the time here involved is not calendar, but 'model-time'. On reflection it is not clear what that means unless it is 'computer time' which is meant and, if it is, one must again ask whether the construction will then have anything to do with any actual price mechanism.

Arrow (1959) has suggested an alternative interpretation which, however, much restricts the applicability of the tâtonnement. Suppose we think of time as divided into trading periods and let the auctioneer follow the rule:

$$p_i(t) = p_i(t-1) + k_i\theta_i[p(t-1)]k_i > 0$$

(with the usual boundary condition to avoid negative prices). Now suppose (*a*) that one is concerned with a pure exchange economy and (*b*) that all goods last for only one period so that agents in each period receive new endowments (identical for each period). Then we can allow the agents to trade during the process *without the trade in any one period* affecting the excess demand at any p in a subsequent period. So now (*a*) we think of the process in real time and (*b*) even if it converges to p^* only as $t \to \infty$ or does not converge at all, agents can trade.

This very restrictive case clarifies the reason why in general the tâtonnement prohibits trade out of equilibrium. Let $\hat{e} = (e^1, \ldots, e^H)$, the endowment matrix of a pure exchange economy in which goods are durable. Let us now take explicit note of $\hat{e}$ in the excess demand function (since it was constant it was omitted hitherto) and write

$$\sum_h \left(x^h - e^h\right) = \hat{\theta}(p,\hat{e}).$$

Assuming that $\theta(p,\hat{e}) = 0$ has a unique solution, the latter will depend on $\hat{e}$ and may be written as $p^*(\hat{e})$. If now trading takes place out of equilibrium, $\hat{e}$ will be changing and so therefore will $p^*(\hat{e})$. Thus when there is such out of equilibrium trading, the equilibrium which the tâtonnement is groping for will depend on the manner of the groping. To exclude this dependence was the purpose of excluding out of equilibrium trade. But there was another reason, namely, the lack of any clear theory of how trade would proceed when either some prospective buyers or sellers could not carry out their trading intentions.

The fictitious auctioneer is also a consequence of theoretical lacunae and indeed of a certain logical difficulty. If prices are to be changed by the economic agents of the theory, that is either by households or firms or both then it is not easy to see how those same agents are also to treat prices as given exogenously as is required by the postulate of perfect competition. This difficulty was first noted by Arrow (1959) who argued that out of equilibrium price changes not brought about by an auctioneer require a departure from the perfect competition assumption if they are to be understood. Take for instance a situation for which $\hat{\theta}_i(p,\hat{e}) > 0$. Then at p there will be unsatisfied buyers. But that means that any firm raising its price for good i by a little will not, as in the usual perfect competition setting, lose all its customers. The reason is that buyers cannot be sure of obtaining the good from any of the other firms which have not yet raised their price. Hence the demand curve for good i facing a producer of that good is not perfectly elastic. (On the other hand, in equilibrium it well might be.) The postulate of the auctioneer sidesteps these problems at the cost of an understanding of how prices are actually changed. It has enabled theorists to ignore the role of monopolistic competition in the process of price formation – a circumstance which until recently has left the whole matter without proper theoretical foundations.

But it must also be admitted that there are formidable theoretical difficulties to be faced in banishing the auctioneer. Whether we think of prices as formed by a bargaining process or by monopolistic competition or in some form of auction process, strategic considerations, that is to say, game theoretic tools, will be required. In addition, careful attention will have to be given to the information available to each of the agents involved in the process. Some progress has been made (e.g. Roth 1979; Schmeidler 1980; Rubinstein and Wolinsky 1985) but there is a very long way to go. (Some economists have banished the auctioneer without considering these matters by the simple device of treating it as axiomatic that at all times the economy is in competitive equilibrium. There is nothing favourable to be said for this move.)

There is now also a somewhat subtler point to consider: the behaviour postulated for the auctioneer will implicitly define what we are to mean by an equilibrium: that state of affairs when the rules tell the auctioneer to leave prices where they are. But the auctioneer's pricing rules are not derived

from any consideration of the rational actions of agents on which the theory is supposed to rest. Thus the equilibrium notion becomes arbitrary and unfounded. If, on the other hand, we had a theory of price formation based on the rational calculations of rational agents then the equilibrium notion would be a natural corollary of such a theory. For instance, one might then be led to describe a situation in which there is unemployment as one of equilibrium because neither firms nor workers, given their information and beliefs, find it advantageous to change the wage.

This line of reasoning leads one to a central objection to the auctioneer and indeed the tâtonnement: it sidesteps the important question of the coordinating power of the price mechanism. Here is an example. In an oligopolistic industry with excess supply it may not be advantageous for any one firm to reduce its price given its beliefs as to the strategies of its competitors. Yet it may be to all of the firms' advantage to have the price reduced: there is a cooperative solution which dominates the competitive one. Put another way, there are significant externalities in price signalling. To leave these unstudied is to leave very important matters in darkness. The auctioneer is a coordinator *deus ex machina* and hides what is central.

These considerations are most striking in the context of Keynesian theory. As long as the auctioneer is in the picture no state of the economy in which there is involuntary unemployment can qualify as an equilibrium – the auctioneer would be reducing wages. But without the auctioneer the observation that a worker would prefer to work at the going real wage to being idle does not logically entail the proposition that the wage will be reduced. That proposition would require a great deal of further theoretical underpinning turning on the beliefs of workers, the strategies of other workers and the strategies of employers. It would also turn on the information available to agents. For instance, if lowering one's wage is regarded as a signal of lower quality of work then one may be reluctant to offer to work at a lower wage. The fictitious auctioneer makes sure that none of these matters is studied or understood. The use of this fiction encourages the view that all Pareto-improving moves will, in a competitive economy, be undertaken. This view, however, lacks any foundations other than the auctioneer himself.

One might just about convince oneself that, notwithstanding all these objections, the tâtonnement and its auctioneer are worthwhile, if it were the case that it provided one story which showed how equilibrium was brought about. Unfortunately, however, it does not do this for there are only a few special cases for which the auctioneer process leads the economy to an equilibrium. In many others it will not do so. Indeed, in so far as one holds the view that an equilibrium is the normal state of an economy one should not be tempted to understand this circumstance by means of a tâtonnement.

See Also

- ▶ Tâtonnement and Recontracting
- ▶ Walras, Léon (1834–1910)

Bibliography

Arrow, K.J. 1959. Towards a theory of price adjustment. In *The allocation of economic resources*, ed. M. Abramovitz et al., 41–51. Stanford: Stanford University Press.

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco/Edinburgh: Holden-Day/Oliver and Boyd.

Heal, G.M. 1973. *The theory of economic planning*. Amsterdam/London: North-Holland.

Roth, A.E. 1979. *Axiomatic models of bargaining*. Lecture Notes in Economics 170. Berlin: Springer.

Rubinstein, A., and A. Wolinsky. 1985. Equilibrium in a market with sequential bargaining. *Econometrica* 53: 1133–1150.

Samuelson, P.A. 1941–1942. The stability of equilibrium. *Econometrica* 9, 97–120; 10, 1–25. Reprinted in Samuelson (1966), vol. 1, 539–562, 565–589.

Samuelson, P.A. 1966. In *The collected scientific papers of Paul A. Samuelson*, ed. J.E. Stiglitz. Cambridge, MA: MIT PressIn.

Schmeidler, D. 1980. Walrasian analysis via strategic outcome functions. *Econometrica* 48: 1585–1593.

Smale, S. 1976. A convergent process of price adjustment and global Newton methods. *Journal of Mathematical Economics* 3 (2): 107–120.

Walras, L. 1874–7. *Eléments d'économie politique pure*. Definitive edn, Lausanne, 1926. Trans. W. Jaffé as *Elements of Pure Economics*. London: George Allen and Unwin, 1954.

Auctions

Vernon L. Smith

Herodotus reports the use of auctions as early as 500 BC in Babylon (see Cassady 1967, pp. 26–40 for references to this and the following historical notes). The Romans made extensive use of auctions in commerce and the Roman emperors Caligula and Aurelius auctioned royal furniture and heirlooms to pay debts. Roman military expeditions were accompanied by traders who bid for the spoils of war auctioned *sub hasta* (under the spear) by soldiers. In AD 193 the Praetorian Guard seized the crown from the emperor Pertinax and auctioned it to the highest bidder, Didius, who, upon paying each guardsman the winning bid, 6250 drachmas, was declared emperor of Rome. It would appear that the Romans used the 'English' progressive method of auctioning, since the word auction is derived from the Latin root *auctus* (an increase).

Types of Auction Institutions

Auctions may be for a single object or unit, as in the unique object auctioning of paintings and antiques at Sotheby's and Christie's in London, or a lot or package of non-identical items, as in the family groups sold in the slave auctions of the antebellum South. Alternatively, auctions may be for multiple units where many units of a homogeneous standardized good are to be sold, such as gold bullion in the auctions conducted by the International Monetary Fund and the US Treasury in the 1970s, and in the weekly auctioning of 91-day and 182-day securities by the Treasury.

Auctions may also be classified according to the different institutional rules governing the exchange. Since the seminal work of Vickrey (1961), it has been recognized that these rules are important because they can affect bidding incentives, and therefore the terms and the efficiency of an exchange. The literature (Cassady 1967; Arthur 1976) has identified many different auction institutions throughout the world, but, following Vickrey (1961), it has become standard to distinguish four primary types of auctions which can be used either in single object or multiple (identical or non-identical) unit auctions.

English Auction

The auction customarily begins with the auctioneer soliciting a first bid for the object from the crowd of would-be buyers, or (where permitted by the auction house rules) announcing the seller's reservation price. Any bid, once recognized by the auctioneer, becomes the standing bid which cannot be withdrawn. Any new bid is admissible if any only if it is higher than the standing bid. The auction ends when the auctioneer is unable to call forth a new higher bid, and the item is 'knocked down' to the last (and highest) bidder at a price equal to the amount bid.

Where multiple units of identical, or nearly identical (close substitute), items are sold by one or more sellers at English auction, individual lots or units are put up for sale in some sequence with each lot or unit sold as a single object. Examples include livestock in the United States and wool in Australia. When there are Q strictly identical items to be sold in a progressive auction, the following alternative procedure has been suggested: '. . . the items are auctioned simultaneously, with up to (Q) bids permitted at any given level, the rule being that once (Q) bids have been made equal to the highest bid, any further bid must be higher than this' (Vickrey 1976, p. 14).

Dutch Auctions

Under this procedure, originally called 'mineing', the price begins at some level thought to be somewhat higher than any buyer is willing to pay, and the auctioneer decreases the price in decrements until the first buyer accepts by shouting 'mine'. The item is then awarded to that buyer at the price accepted. Many years ago this procedure was automated by an electrical clock mechanism which is used widely in Holland for the sale of produce and cut flowers. The clock is normally located in a large amphitheatre (Cassady 1967, p. 194) with buyers sitting at desks facing the clock. An indicator hand on the clock decreases counterclockwise through a series of descending

prices. Any buyer can stop the indicator hand by pressing a button when the descending indicated price is acceptable.

The descending offer procedure is used in the sale of fish in England and Israel, in the sale of tobacco in Canada, and a variant of the procedure is used regularly to mark down clothing in Filene's department store in Boston. When the descending offer procedure is applied to multiple units, the first bidder exercises his option to take any part of the quantity offered. The offer price then continues its descent until the next bidder accepts and so on. Thus in fish markets in British ports, the auctioneer accepts 'book' bids for specified quantities. If the offer price reached this level before anyone accepts from the floor, the book bid is filled with any remaining quantity offered at descending prices to the crowd.

First Price Auction

This is the common form of 'sealed' or written bid auction, in which the highest bidder is awarded the item at a price equal to the amount bid. The multiple unit generalization of this procedure is called a *discriminative* auction. Thus if Q identical units are offered, the highest bids for the first Q units are all accepted at the prices and quantities stated in the bids tendered. The weekly primary auction of new short-term US. Treasury securities has used this institution for about fifty years.

Second Price Auction

This is a sealed bid auction in which the highest bidder is awarded the item at a price equal to the bid of the second highest bidder. The procedure is not common although it is used in stamp auctions. For example, the London stamp auction uses English oral bidding, but buyers not present may submit written 'book' bids. An award to a book bidder is made at one price interval or unit above the floor bid, or the second highest book bid, whichever is the largest. If the auctioneer has two book bids he starts the bidding at a unit interval above the second highest of the book bids. If the bid is not raised on the floor he declares it sold to the highest book bidder at this (approximately) second highest bid price.

The multiple unit extension of the second price sealed bid auction is called a *competitive* (or uniform price auction). Under this procedure if Q identical units of a good are offered, the highest bids for the first Q units are all accepted at one market clearing price equal to the bid for the Q + 1st unit. The procedure was used experimentally by the US Treasury in the 1970s to sell long-term bonds, and in one gold bullion sale. Exxon corporation has sold bonds (usually to registered brokers and dealers) by this method on several occasions since the US Treasury experiments. Since 1978 Citicorp has been auctioning commercial paper weekly using the method, but the institution has not found general acceptance. These auctions are referred to as 'Dutch' auctions in the financial trade literature, but this is a misnomer because the long established 'mineing' procedure, known as the Dutch auction, follows a discriminative, not a uniform, multiple unit pricing procedure.

A summary of auction institutions should not omit some comment on the Walrasian *tâtonnement* hypothesis, which has long served the need of equilibrium price theory for a path independent process that precludes contracting at non-equilibrium prices. It appears that the only naturally occurring organized markets using a procedure similar to a Walrasian *tâtonnement* are the gold and silver bullion price 'fixing', or determining, markets (Jarecki 1976). In the London Gold Market, representatives of the dealers in this market meet twice daily, and establish a price as follows: the chairman of the meeting begins with an initial starting price, and each representative indicates whether he is a seller, a buyer or neither at that price. Each dealer has orders from clients all over the world. To be a buyer means that at the trial price and volume of his client's buy orders exceed the volume of the sell orders. If at the starting price there are no sellers, the price is raised by varying amounts until one or more of the traders indicates that he is a seller at the standing price. Similarly, the price is moved down if there are no buyers at the starting price. At this juncture the chairman asks for 'figures'; i.e. for the net quantities each trader

wishes to buy or sell. If the total indicated purchase quantity does not match the quantity offered by the traders the price is further adjusted until a match occurs. This Walrasian market also has a unanimity stopping rule. Each trader has a small Union Jack in front of him. When a trader is satisfied with the standing price, and has no further orders that require price adjustment, he puts the flag down. The chairman announces that the price is 'fixed' if and only if all flags are down.

Theory of Auctions

The following analysis of auctions will adopt five principal assumptions: (1) Each bidder desires to purchase a single unit of the commodity. (2) Buyer i associates a cash value, v_i, with the item which represents i's maximum willingness to pay. In some auctions, notably the English institution, v_i can be interpreted as the cash equivalent of an uncertain item value. (3) The value v_i to i is independent of the value, v_j to any j; i.e. v_i would not change if i had knowledge of v_j for all i and j. (4) Each i knows the value v_i, but has no certain knowledge of the values of others. (5) Transactions costs, including the cost of thinking, calculating, deciding and bidding, are negligible. Without loss of generality we can number the agents so that $v_1 > v_2 > \ldots > v_N$. An auction allocation is efficient (Pareto optimal) if it awards the offered unit(s) to the buyer(s) that value it most highly. When $Q = 1$ unit is offered, the allocation is efficient if it goes to buyer 1 with value v_1. If $Q = 7$ is offered, an efficient allocation requires buyers 1 to 7 each to receive one unit.

The English and Dutch systems are *continuous auctions* (in time) in which an agent may alter his/her bid in response to the bids of others, or the failure of a bid to be accepted; i.e. bid information is made available continuously by the process until the auction stopping rule is invoked. In *sealed bid* auctions each agent submits one bid message to a centre, which processes the messages according to the rules of the institution, then announces some form of aggregate or summary information describing the outcome. Either type of auction may be repeated over time, thereby generating a history of outcome information, but continuous auctions provide a message history between successive contracts, while sealed-bid auctions do not.

In auction theory it is convenient to define formal concepts of environment, institution, and agent behaviour (*see* ▶ Experimental Methods in Economics). The *environment,* $E = (E^1,\ldots, E^N)$, where each agent's characteristics, $E^i = (u^i, w^i, T^i)$, are defined by his preferences or utility (u^i), endowment (w^i), and state of knowledge (T^i). In the English or Second Price auction, $E^i = (v_i, N > 1)$ for agent i, indicating that i's preferences and endowment are defined by his/her value for one unit of the commodity, that i knows that there is at least one other bidder, and (by omission) that i knows nothing about any $v_j, j \neq i$.

The institution specifies (1) a language, $M = (M^1,\ldots,M^N)$, consisting of message elements $m = (m^1,\ldots,m^N)$, where M^i is the set of messages that can be sent by i, and m^i is the message sent by i; (2) a set of allocation rules $h = [h^1(m),\ldots, h^N(m)]$, and a set of cost imputation rules $c = [c^1(m),\ldots,c^N(m)]$, where $h^i(m)$ is the commodity allocation to agent i, and $c^i(m)$ is the payment required of i, given all the messages, m; (3) a set of adjustment process rules, $g(t_0, t, T)$, consisting of a starting rule, $g(t_0,\cdot,\cdot)$, a transition rule, $g(\cdot,t,\cdot)$, and a stopping rule, $g(\cdot,\cdot,T)$, after which the allocation and cost imputation rule become effective. Hence, an *institution* is defined by $I = (I^1,\ldots, I^N)$, where $I^i = [M^i, h^i(m), c^i(m), g(t_0, t, T)]$. In all auctions the messages are bids; i.e. $m^i \equiv b_i$, where b_i is a bid by agent i. Let the bids be numbered from highest to lowest $b_1 > b_2 > \ldots > b_N$ (the order and numbering of the bids need not be the same as for the values). In an English auction the process starts with some bid $b_j(t_0)$ by some agent j. This is the standing bid until, under the transition rule, some agent announces a higher bid which becomes the new standing bid, and so on in sequence. The process stops with a bid $b_1(T)$ when the auctioneer is unable to solicit a higher bid. Hence, $b_1(T)$ becomes the final message, and in the English auction institution, $I_e = \left(I_e^1,\ldots,I_e^N\right)$, the outcome rules are

$$I_{\mathrm{e}} = \left[\cdot, h^1(m) = 1, c^1(m) = b_1; \right.$$
$$\left. h^i(m) = 0, c^i(m) = 0, \text{ for all } i > 1, \cdot\right]$$

indicating that the last (and highest) bidder wins the item, pays the amount bid, and all others receive and pay nothing. In the Second Price sealed-bid auction the starting and stopping rules merely define the pre-auction time interval within which bids are to be tendered, and there is no transition rule. The bids are all examined at once, and the Second Price institution specifies $I_s = [h^1\ (m) = 1,\ c^1\ (m) = b_2;\ h^i\ (m) = 0,$ $c^i\ (m) = 0$, for all $i > 1$], indicating that the high bidder is awarded the item at a price equal to the next highest bid, with all others receiving and paying nothing.

Within this framework we define *agent behaviour* as a function that carries each agent's characteristics, E^i, given the institution, I, into the (final) message m^i sent by i, $m^i = \beta(E^i \mid I)$. A *theory* of agent behaviour has the objective of specifying β as a *hypothesis* about the observed message responses of agents in alternative institutions such as the English and Second Price auctions.

English

Let $b_k(t)$ be the t th standing bid (in some sequence), announced by agent k. Then it is a dominant strategy for any $i \neq k$ to raise the bid if $v_i > b_k\ (t)$; i.e. this strategy is *best* for i whatever might be the response of any other agent. Note that since the winning bidder must pay the amount bid it is never optimal for any i to raise her own bid. If the auction has a standard bid increment, δ assumed to be smaller than the distance between any two adjacent values, then i is motivated to bid $b_i(t+1) = b_k(t) + \delta$ if and only if $v_i \geqq b_i(t+1)$. Clearly, this process must stop with the Tth bid, when (eventually) agent 1 bids $b_1\ (T)$, where $v_2 - \delta < b_1\ (T) \leqq + v_2 + \delta$, and agent 2 is unable to raise the bid without bidding in excess of v_2. Hence, in the English auction we have $m^i \equiv b_i = \beta(v_i,\ N > 1 |\ I_e) \equiv v_i$ for $i = 2,\ 3,\ \ldots,\ N$, i.e. each $i \neq 1$ is motivated to reveal demand by bidding up to his value v_i, with agent 1 *discovering* that she does not need to bid v_1, but at most $v_2 + \delta$ to obtain the award. It follows that the equilibrium price, p_{e}, must satisfy $v_2 - \delta < p_e \leq v_2 + \delta$, and the award to agent 1 will be efficient.

Because individual units are sold sequentially in typical multiple unit English auctions ($N > Q > 1$), a theory of this case would require some hypothesized expansion of agent information sets which allows each i to weigh formally the prospect of underbidding v_i by some amount in earlier auctions in anticipation of possible lower prices in later auctions. But Vickrey's generalization (quoted above) of the English auction to multiple identical units, which preserves the information properties of the single unit case, does lead to determinate results: once the bidders with the Q highest values match bids at $b(T) \in (v_{Q+1} - \delta,$ $v_{Q+1} + \delta)$, then no bidder will be motivated to raise this standing bid. Hence, the price for any Q units ($N > Q \geq 1$) must satisfy $v_{Q+1} - \delta < p_{\mathrm{e}}$ $< v_{Q+1} + \delta$, and the award to agents $1, 2, \ldots Q$ will be efficient.

Second Price

In this auction the surplus obtained by the winning bidder depends upon the bid of the highest among the other $N - 1$ losing bidders; i.e. if i is the winner and j the highest losing bid, the surplus to i is $v_i - b_j$. Hence the optimal bid is the bid that maximizes the probability of winning a positive surplus. This occurs only if each i bids v_i. To bid less than v_i is to reduce the chance of being the high bidder, without affecting the surplus $v_i - b_j$. To bid more than v_i is to risk (without compensating benefit) winning at a price $b_j > v_i$, yielding a negative surplus. If each i *reasons* in this manner, then $m^i \equiv b_i = \beta(v_i,\ N > 1 |\ I_s) \equiv v_i$ for *all i*. It follows that the award will be to agent 1, which is efficient, and the price will be $p_s = b_2 = v_2$. This argument extends to the multiple unit case in which N bidders each submit a bid for one of Q identical units ($N > Q > 1$). It is a dominant strategy for each i to bid v_i, the award will be to agents $1, 2, \ldots Q$, and the competitive price paid by all Q winning bidders will be $p_c = b_{Q+1} = v_{Q+1}$.

In comparing the English and Second Price institutions it is seen that in the limit, as δ becomes small, the two institutions are isomorphic; that is, they lead to the same price and allocations. In the language of game theory these institutions are

equivalent in the sense that they have the same normal form. They have quite different extensive (sequential process) forms. Analysis of the richer extensive form of the English auction leads to the conclusion that the high bidder wins with a bid of v_2 which makes the theoretical auction outcome identical to that of the Second Price auction, although the institutions have distinct cost imputation rules. It should be noted from our discussion in section I that the Second Price procedure appears to have arisen in practice in the British stamp (and some fish) markets which permitted 'book' bids at English auction. It is easy to see that in such circumstances auctioneers might soon 'discover' the equivalence of the English and Second Price procedure without having to resort to formal analysis.

The First Price and Dutch auctions use the same allocation and cost imputation rules that are used in the English auction; they are like the Second Price auction in that the auction is over before the bidders obtain informative data about their rivals from the auction itself. In these auctions it is of importance what each bidder assumes about the values and bidding behaviour of his rivals. In the analysis below we will follow Vickrey (1961) in supposing that the values are assumed by each agent to be independent occurrences from a constant density on the interval [0, 1]. Any bids and values can be mapped into this interval by expressing them as fractions of the largest possible value. Thus, if the maximum value is $\overline{v}$, a bid of b' and value v', can be represented by $b = (b'/\overline{v})$ and $v = (v'/\overline{v})$. With these assumptions about agent knowledge the environment is $E^i = [v_i; P(v) = v, N > 1]$ indicating that each i knows with certainty his/her own value $v_i \in [0,1]$ for a single unit, that the other agent's values have the probability distribution, Prob $\{X < v\} \equiv P(v) = v$ and that there are N bidders.

First Price

Vickrey (1961) showed that if all agents are risk neutral the noncooperative (or Nash) equilibrium bid function in the First Price auction is

$$m^i \equiv b_i = \beta\left[v_i; P(v) = v, N | I_f\right] \equiv \left(\frac{N-1}{N}\right) v_i.$$

If all bidders have the same strictly concave utility function for surplus, say $u(v_i - b_i)$, the resulting bid function $b_A(v_i)$ will have the property

$$b_A(v_i) > \left(\frac{N-1}{N}\right) v_i$$

(Holt 1980). In both of these cases, since the equilibrium bid function depends only on value, and not upon which agent has any particular value, any given ordering of the values induces the same ordering on the bids. Hence the highest value bidder will submit the highest bid, and the allocation is efficient. However, if each bidder i has constant relative risk averse (CRRA) utility, $(v_i - b_i)^r i$, $r_i \in (0,1)$, then it can be shown (Cox et al. 1982) that the Vickrey bid function generalizes to

$$b_i = \left(\frac{N-1}{N-1+r_i}\right) v_i, \quad \text{for } b_i < \overline{b} = \frac{N-1}{N}.$$

Consequently, in this case (and in general when utility functions are distinct) the highest value bidder is not necessarily the highest bidder, since if he is less risk averse than the second, or third, highest bidder, his bid may be lower than theirs. All these results have been further generalized to the multiple unit discriminative auction $(N > Q > 1)$ (see Vickrey 1962; Harris and Raviv 1981; Cox et al. 1984).

Dutch

The Dutch auction starting rule is to announce (or display on the clock) an initial asking price, $a(t_0)$. If the clock speed, measured in dollars, is s ($ per second), then the transition rule states that at time t the asking price is $a(t) = a(t_0) - st$ If at T, agent i is the first to accept the standing offer (the stopping rule), then i's bid, and the price paid, is $b_i = a(t_0) - sT$. Each bidder must decide when to stop the descending offer price. Vickrey was the first to argue that the Dutch and First Price auctions are isomorphic; i.e. that a bidder i who would bid b_i in the First Price auction would stop the clock at T such that $b_i = a(t_0) - sT$ in the Dutch auction. This was demonstrated

formally in Cox et al. (1982) by proving the equivalence of a pre-auction planning model of the Dutch (and First Price) auction with a Bayesian model of participation in the Dutch auction. The Bayesian model shows that the information at time t on the Dutch clock (no bidder by time t has stopped the clock) is non-informative; i.e. it provides no rational basis for modifying the optimal bid given any pre-auction postulated environment, such as $E_i = [v_i,\ P(v),\ N > 1]$. Hence, a Nash model of the Dutch auction (assuming CRRA utility) yields the behavioural hypothesis that

$$m^i \equiv b_i = \beta[v_i; r_i; P(v) = v, N > 1 | I_d] \times \equiv \left(\frac{N-1}{N-1+r_i}\right) v_i,$$

where each i is defined by the characteristics (v_i, r_i).

Because the Dutch auction has such a rich extensive form, containing parameters such as $a(t_0)$ and s that do not enter into the First Price auction, it would be surprising if these two auction procedures produced the same results in any particular parametric implementation. One can easily imagine an s so large that the standing price is not discernible on a Dutch clock, with a bidder having to guess at the bid price at which she is stopping the clock. Similarly, s might be so small that the waiting cost is significant leading to higher bids in the Dutch than in the First Price auction. The Dutch–First Price equivalence theorem abstracts from these extensive form parametric differences and analyses each institution as a mathematical game in normal form.

Theoretical behaviour in the standard single object auctions can be compared using the following compact representation:

$$b_i = \beta(E^i|I) = \begin{cases} b_1 \in (v_2 - \delta,\ v_2 + \delta) \text{ for } i = 1, \text{ and } b_i \leqq v_i, \text{ for } i > 1, \\ \text{if } E^i = (v_i),\ I = I_e \\ v_i, \text{ for all } i, \text{ if } E^i = (v_i; N > 1),\ I = I_s \\ \left(\frac{N-1}{N-1+r_i}\right) v_i, \text{ for all } i, \\ \text{if } E^i = [v_i, r_i; \Phi(r),\ P(v) = v, N > 1],\ I = I_f \text{ or } I_d \end{cases} \tag{1}$$

These results generalize for multiple units, giving

$$b_i = \beta(E^i|I) = \begin{cases} b_Q \in (v_{Q+1} - \delta,\ v_{Q+1} + \delta) \text{ for } i \leqq Q, \\ \text{and } b_i \leqq v_i \text{ for } i > Q, \text{ if } E^i = (v_i), I = I_e \\ v_i, \text{ for all } i, \text{ if } E^i = (v_i; N > Q \geqq 1), \\ I = I_{Q+1} (Q+1 \text{ price auction}) \\ b_a(v_i, r_i | E(r), N > Q \geqq 1), \\ \text{if } E^i = [v_i, r_i; \Phi(r),\ P(v) = v, N > Q \geqq 1], \\ I = I_D (\text{discriminative auction}), \end{cases} \tag{2}$$

where b_a is the CRRA bid function for multiple units [see Cox et al. (1984) and the references therein for the formula and its derivation], φ (r) is the population distribution of the CRRA risk parameter and $E(r)$ is its expected value.

If we let $E[P(I)]$ be the (mathematical) expected selling price in a single object auction under institution I, using (1) it is easy to compare the four standard auctions in terms of this outcome measure (Vickrey 1961) if we assume risk neutrality; i.e. E^i $(v_i,\ r_i = 1;\ P(v) = v,\ N > 1)$:

$$E[P(I)] = \begin{cases} E(b_1) \in \left[\left(\frac{N-1}{N+1}\right) - \delta, \left(\frac{N-1}{N+1}\right) + \delta\right], \text{ if } I = I_e, \\ E(b_2) = \left(\frac{N-1}{N+1}\right), \text{ if } I = I_s, \\ E(b_1) = \left(\frac{N-1}{N+1}\right), \text{ if } I = I_f \text{ or } I_d, \end{cases} \tag{3}$$

since $E\,(v_2) = (N - 1)/(N + 1)$ It follows that for $\delta = 0$, all four auctions give the same expected selling price. It is also easy to show that if the bidders are risk averse, then $E[P(I_e)] = E[P(I_s)] < E\left[P\left(I_f\right)\right] = E[P(I_d)]$. Thus a testable outcome implication of the above models of bidding behaviour is that observed mean prices will be ordered.

$$\overline{P}(I_e) = \overline{P}(I_s) = \left(\frac{N-1}{N+1}\right) \leqq \overline{P}\left(I_f\right) = \overline{P}(I_d).$$

Also, efficiency, measured by the percentage (probability) of awards to the highest value bidder will be 100 per cent in all the auctions if bidders are

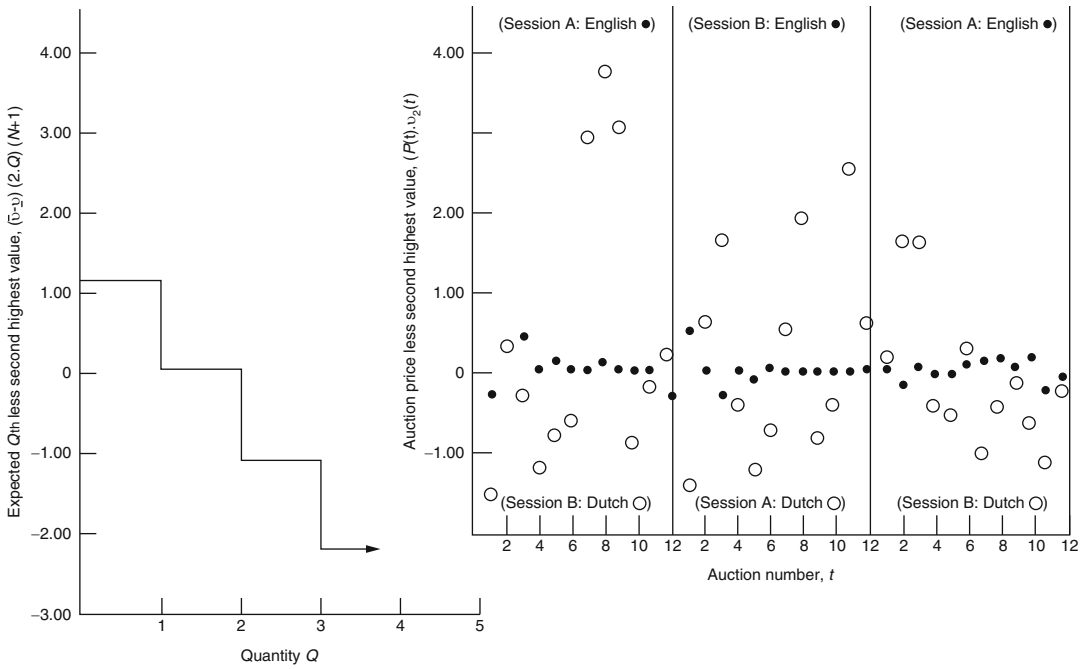
risk neutral or all have the same concave utility for surplus. But if bidders have CRRA utility with different parameters, r_i then this measure of efficiency, ζ, will be ordered $100 = \zeta_e = \zeta_s > \zeta_f = \zeta_d$ in the English, Second, First and Dutch auctions respectively.

Experimental Tests of Auction Market Behaviour

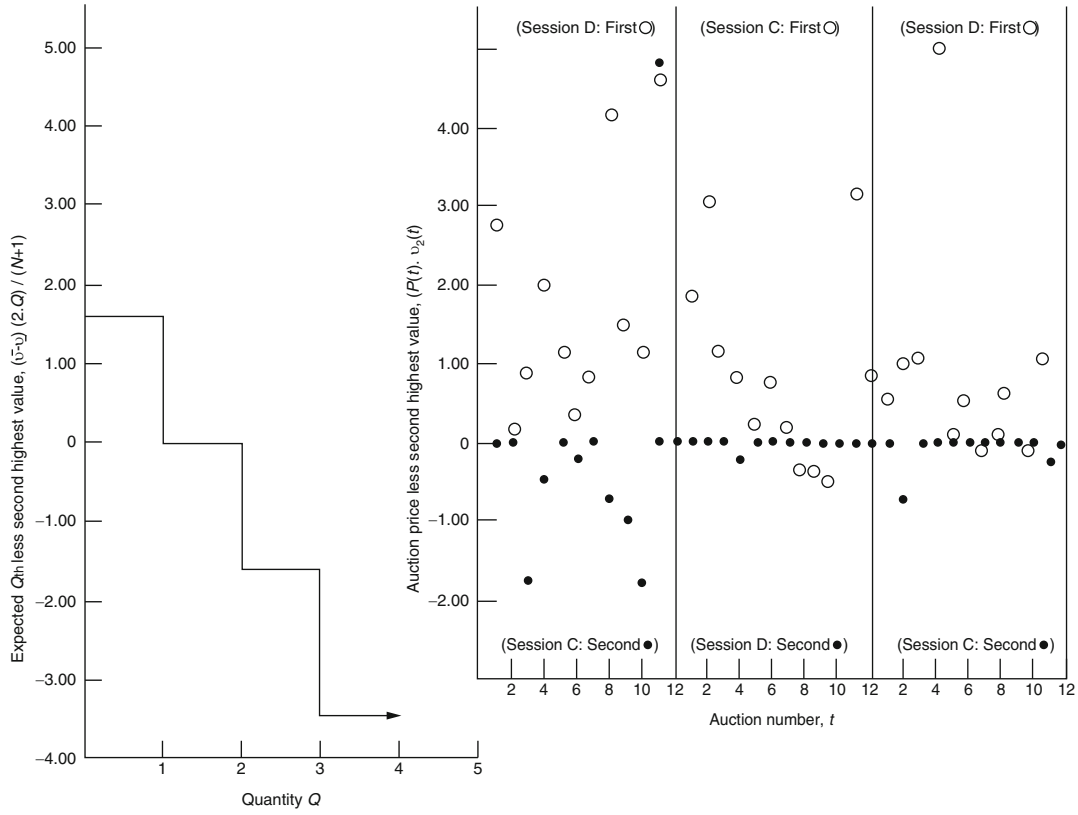
Several studies (see Cox et al. 1984 and its citations) have tested the above models and various extensions of them using experimental methods. In all of the experiments summarized below, values are assigned from a uniform probability function whose parameters are common knowledge to all the participants. Each participant understands that he/she will be paid in cash the difference between the value assigned, and the price paid, conditional upon being a winning bidder in any particular auction.

From the numerous experimental studies reporting the results of perhaps 1500–2000 auctions, the following brief summary is offered:

1. The behaviour of prices in the four standard auctions is illustrated by the representative charts in Figs. 1 and 2 comparing I_e and I_d prices using eight bidders, and I_f and I_s prices using five bidders. In these experiments, in each auction, a random sample of N values are assigned to the bidders from the uniform distribution on the interval $[\underline{v}, \overline{v}]$. From the distribution function (order statistic) for the Qth highest value in a sample of size N one can compute the expected Qth highest value, which is the expected (Marshallian) demand schedule, $E(v_Q|N) = [(\overline{v} - \underline{v})(N - Q + 1)/(N + 1)] + \underline{v}$. This schedule is graphed on the left of Figs. 1 and 2 in normalized form by subtracting the expected second highest value; i.e. $E(v_Q|N) = (\overline{v} - \underline{v})(2 - Q)/(N + 1)$ is graphed. Similarly, on the right of Figs. 1 and 2 are plotted the prices realized in each auction, normalized by subtracting the second highest value realized in each sample. Normalized in this way the risk neutral predicted average price is zero in all auctions. Figure 1 charts the prices in 36 sequential auctions in each of two experimental sessions with



Auctions, Fig. 1 English-Dutch prices compared: eight bidders $[\overline{v}, \underline{v}] = [\$6.1, \$10]$



Auctions, Fig. 2 First-Second prices compared: five bidders $[\overline{v}, \underline{v}] = [\$0.1, \$10]$

different groups of size $N = 8$. Session A consisted of 12 English, followed by 12 Dutch, and ending with 12 English, auctions. Session B consisted of the opposite Dutch – English – Dutch sequence of 12 auctions each. A similar paired comparison of First and Second auctions using five bidders is shown in Fig. 2. These charts should help to disabuse anyone of the notion that pricing institutions do not matter.

2. The mean observed prices in the four standard auctions (conducted under conditions which control for N, and other such parameters in the comparisons) for $N = 4, 5, 6, 8$ and 9 satisfy the ordering $\overline{P}(I_e) \cong E(v_2) \cong \overline{P}(I_s) < \overline{P}(I_d) < \overline{P}(I_f)$. Actually, mean prices in I_s tend to be below those in I_e because many subjects initially do not follow the dominant strategy rule $b_i = v_i$, but over time more and more subjects 'learn' to adopt this strategy. An example is shown in session C of Fig. 2, in which six of the twelve prices in the First sequence of 12 auctions under *Is* are below v_2, but in the second sequence (last panel) under I_s only two of the 12 prices are below v_2. Taking account of this convergence over time we can say that observed English and Second prices support the price implications of the theory as stated in (3).
3. Efficiences in the English and Second auctions are approximately the same (97 per cent and 94 per cent respectively) but are much lower in the First auction (88 per cent) and still lower in the Dutch (80 per cent).
4. These price and efficiency results support the following conclusions: (a) the English and Second auction are approximately equivalent;

(b) the Dutch and First Price auctions are not isomorphic, behaviourally; (c) the First auction results are consistent with risk averse Nash equilibrium behaviour; and (d) both the efficiency data and observations on individual bidding support the CRRA model of Nash equilibrium bidding, with different bidders exhibiting different degrees of risk aversion in their bidding behaviour; (e) the CRRA model of bidding in the First auction is also supported by the finding that increasing the payoff levels by a factor of three [paying the winning bidder 3 ($v_i - b_i$) instead of ($v_i - b_i$) dollars] has no effect on bidding behaviour – a theoretical result which follows if and only if utility is of CRRA form.

5. An extensive study of multiple unit discriminative auctions finds that the data are consistent with the CRRA Nash model of bidding behaviour over much but not all of the (N,Q) parameter space (Cox et al. 1984). Hence, anomalies remain, and in view of the highly replicable and non-artifactual character of the empirical results there is the strong implication that the resolution of these anomalies is an unfinished theoretical task.
6. The Second Price auction results do not extend to the multiple unit uniform price auction. Apparently, with multiple units, in those parameter cases that have been studied, the market is less effective in disciplining (with failure experiences) those strategies that depart from the dominant strategy.

Enriching the Environment: Dependence, Information, Collusion and Combinatorial Considerations

Once replicable experimental results have been established and the strengths and weaknesses of a theory have been assessed, it is natural to extend both the theoretical and the empirical inquiry to richer environments. The required theoretical advances have been more difficult to achieve than the creation of richer environments in the laboratory (Kagel et al. 1983).

A limiting feature of the above theories is the assumption that agent values are independent. Consequently, each agent's willingness-to-pay as might be revealed in the open English auction has no information value to any other agent. Milgrom and Weber (1982) capture this important postulated property of some commodities by introducing the concept of positively dependent (affiliated) values. In this environment the English auction is no longer isomorphic to the Second Price auction; instead, prices in the former exceed those in the latter. Milgrom and Weber (1982) also argue that the Dutch–First isomorphism continues to hold when values are positively dependent, but this extension is of more limited scientific significance (than the extension of English–Second auction theory) since the experimental evidence is inconsistent with this implication in the independent values environment. Any theoretical implication found to be robust with respect to some generalization of the environment is a moot discovery if that implication is contrary to the evidence in the more special environment.

An important application of the case in which individual values are affiliated is to the sealed-bid auctions of oil exploration and development leases by the government. In this case we can think of each v_i as i's estimate of the value of the lease after obtaining seismic and other sample data providing information on the existence of possible oil bearing geological structures. This application is often referred to as the common value of mineral rights model, since the analysis has assumed that all companies place the same value on any petroleum that might be discovered on the tract. This assumption is much too limiting since there is an active market for existing or proven petroleum reserves, and one cannot account for such exchanges if private values are indeed common. Hence petroleum exploration and development leases are best viewed as a case in which the commodity exhibits differing, but affiliated, private values. The first experimental study of 'common value' auctions (Kagel et al. 1983) reports bidding behaviour in which the bids are too high to be consistent with risk neutral utility functions and too diverse among

individuals to be consistent with the implication of symmetrical bid functions. In effect these results, when values are affiliated, imply rejection of the common values model, and serve to establish the robustness of the experimental results when values are independent (Cox et al. 1984). These findings heavily underline the methodological point that evidence contrary to a postulate (e.g. symmetry) in any environment requires modification of the theory if one is to obtain empirically useful and observationally disciplined extensions of the theory to more complex environments.

The different standard auctions are not equivalent in terms of their collusive potential. The open-bid English auction is particularly vulnerable to collusion since a subset of $n < N$ buyers have only to agree not to bid against each other in order to reduce the expected price that will be paid. Furthermore, the English auction process assures that the agreement will be easy to monitor.

It is an open question whether a seller, such as the government in the sale of mineral leases, should publicize each bidder's bid every time a sealed-bid auction is conducted. It serves to reinforce the credibility of the auction process by allowing each bidder to verify that his bid was processed honestly. But if a buyers' ring is operating, such information makes it easy for the ring to monitor the bids of its members, and to determine the identity of outside bidders and the conditions under which the ring loses the auction.

Sealed bid auctions are vulnerable to collusion between the auctioneer and one or more buyers, and between the auctioneer and the seller. Thus, in the First Price auction, the terms of agreement between the auctioneer and a buyer might be that if the buyer enters the highest bid, then his bid is to be reentered slightly in excess of the second highest bid.

The Dutch auction is perhaps effective against all of the above examples of collusion. In this auction, since none of the losing bids is known to anyone, they cannot even be leaked, let alone announced, and these types of conspiracies are not feasible.

A new proposed auction institution which has yet to be implemented in practice, but has been subjected to limited testing in the laboratory is the combinatorial auction (Rassenti et al. 1982). This is a sealed-bid auction which allows bidders to submit bids for one or more combinations of non-identical items in a multiple unit auction. The problem was originally suggested in the context of designing a market for airport landing or takeoff slots. Airport slots are an extreme example of a resource whose productive value is enhanced in specified combinations. Thus a slot at New York's Kennedy International has no productive value except in combination with a flight compatible slot at Chicago's O'Hare Field.

See Also

- ▶ Bidding
- ▶ Exchange
- ▶ Experimental Methods in Economics

Bibliography

Amihud, Y. (ed.). 1976. *Bidding and auctioning for procurement and allocation*. New York: New York University Press.

Arthur, H. 1976. The structure and use of auctions. In Amihud (1976).

Cassady, R. 1967. *Auctions and auctioneering*. Berkeley: University of California Press.

Cox, J., B. Roberson, and V. Smith. 1982. Theory and behavior of single unit auctions. In *Research in experimental economics*, vol. 2. Greenwich: JAI Press.

Cox, J., V. Smith, and J. Walker. 1984. Theory and behavior of multiple unit discriminative auctions. *Journal of Finance* 39(4): 983–1010.

Harris, M., and A. Raviv. 1981. Allocation mechanisms and the design of auctions. *Econometrica* 49(6): 1477–1499.

Holt, C. 1980. Competitive bidding for contracts under alternative auction procedures. *Journal of Political Economy* 88(3): 433–445.

Jarecki, H. 1976. Bullion dealing, commodity exchange trading and the London gold fixing. In Amihud (1976).

Kagel, J., J. Levin, R. Battalio, and D. Meyer. 1983. Common value auctions: Some initial experimental results. University of Houston Working Paper, November.

Milgrom, P., and R. Weber. 1982. A theory of auctions and competitive bidding. *Econometrica* 50(5): 1089–1122.

Rassenti, S., V. Smith, and R. Bulfin. 1982. A combinatorial auction mechanism for airport time slot allocation. *Bell Journal of Economics* 13(2): 402–417.

Vickrey, W. 1961. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* 16(1): 8–37.
Vickrey, W. 1962. Auctions and bidding games. In *Recent advances in game theory, proceedings of a conference*. Princeton: Princeton University Press.
Vickrey, W. 1976. Auctions, markets and optimal allocation. In Amihud (1976).

Auctions (Applications)

Patrick Bajari

Abstract

We survey some recent empirical work concerning the analysis of auctions. We begin by describing a two-step nonparametric approach for estimating bidding models that is commonly used in the applied literature. Two applications of this approach are considered: empirical work on bidding in Treasury markets, and empirical tests for collusion in auctions.

Keywords

Auctions; Bid rigging; Cartels; Collusion; Conditional independence; Discriminatory auctions; Exchangeability; First-price auctions; Sealed-bid auctions; Statistical decision theory; Structural estimation; Uniform price auctions

JEL Classifications

D4

In this article, we survey some recently developed methods for the econometric analysis of action data and related applications. Since the mid-1990s, auctions have been an active area of research in empirical industrial organization. Auctions are an attractive setting for empirically testing game theory, for three reasons. First, real-world auctions have well-defined rules, which often correspond closely to game forms in economic theory. The mapping between the data and economic theory is typically less ambiguous in auctions than in other applications in empirical industrial organization. Second, the theoretical literature on auctions is well developed and offers many testable implications. Third, there are many high quality, easily accessible data sets. For example, detailed data sets from public sector procurements or online auctions can easily be collected from the Internet.

In this survey, we shall describe the estimation strategy proposed in Guerre, Perrigne and Vuong (2000) (henceforth GPV) and two substantive applications. The empirical literature in auctions is diverse. Numerous useful alternative approaches have been proposed, so it is impossible to cover all of them in a short survey. However, the work of GPV and related extensions is widely viewed as one of the most important recent additions to the literature. This survey will omit many of the technical details which are required to correctly implement these estimators. Instead, we discuss the estimators somewhat informally, focusing on what we believe is the key intuition behind these methods. Fortunately, there are several excellent surveys that discuss these estimators and related applications in considerable detail. See, in particular, Athey and Haile (2007), Hendricks and Porter (2007) and Hong and Paarsch (2006).

The First-Price Auction

Following GPV, consider a first-price sealed-bid auction with independent private values. There are $i = 1,\ldots,N$ bidders. Bidder i's valuation for winning the auction is denoted by v_i and is private information. The bidders are symmetric in the sense that each bidder's valuation is an i.i.d. draw from a distribution $F(v)$, which is common knowledge. After learning their valuations, each bidder independently and simultaneously submits a bid b_i. Bidders are risk neutral, and bidder i receives utility $v_i - b_i$ if i is the high bidder and zero otherwise. The equilibrium bid function is symmetric and strictly increasing under fairly mild regularity conditions. Let $b = \mathbf{b}(v)$ denote the equilibrium bid function and $\phi(b) = \mathbf{b}^{-1}(v)$ denote the inverse bid function.

Bidder i's expected utility from bidding b_i is equal to

$$(v_i - b_i)F(\phi(b_i))^{N-1}. \quad (1)$$

Bidder i wins the auction when the other $N-1$ bidders bid less than b_i. Bidder $j \neq i$ bids less than b_i when j's valuation is less than $\phi(b_i)$. The probability of this event is $F(\phi(b_i))$. Therefore the probability that bidders $j \neq i$ bid less than b_i is $F(\phi(b_i)^{N-1}$ Expected utility is the product of the surplus bidder i receives conditional on winning, $(v_i - b_i)$, times the probability that i wins the auction. Given v_i, the first- order condition for utility maximization is

$$(v_i - b_i)(N-1)f(\phi(b_i))\phi'(b_i) - F(\phi(b_i) = 0. \quad (2)$$

Suppose that the econometrician observes $t = 1,\ldots,T$ independent repetitions of the auction described above. For each auction t, the econometrician observes all of the bids $b_{i,t}$. The object that GPV wish to estimate is the distribution of bidder valuations, $F(v)$. GPV's approach is structural in the sense that they attempt to recover the economic primitives of the model. As we shall discuss in our applications, structural estimation of the model may allow the economist to answer a number of substantive questions. For example, we can assess the efficiency of the observed auction mechanism or test between competing models, such as competition versus collusion.

GPV note that an econometric approach based directly on evaluating Eq. (2) may be difficult. This equation involves the inverse bid function, ϕ, and its derivative, ϕ', which in turn are complicated, nonlinear functions of the unknown $F(v)$. In principle, it is possible to estimate parametric auction models based on Eq. (2), as in Paarsch (1992), Donald and Paarsch (1993), Hong and Shum (2002) and Bajari and Hortaçsu (2003). However, these methods rely on restricting attention to carefully chosen parametric distributions or require the use of reasonably sophisticated numerical methods. (Despite these limitations, it is worth noting that many parametric approaches generate superconsistent estimators, which converge much more quickly than the nonparametric rate of convergence as in GPV. This may be useful when the sample size available to the econometrician is limited. See Donald and Paarsch 1993; Hirano and Porter 2003, for a discussion.)

A key insight of GPV is that the econometric analysis of the first-price auction is greatly simplified by a change of variables. Let $G(b) = F(\phi(b_i))$ denote the equilibrium distribution of the bids. If we substitute $G(b)$ into (1), we can write expected utility as

$$(v_i - b_i)G(b_i)^{N-1}.$$

The first-order conditions now become

$$(v_i - b_i)(N-1)g(b_i) - G(b_i) = 0 \quad (3)$$

$$v_i = b_i + \frac{G(b_i)}{(N-1)g(b_i)}. \quad (4)$$

The right-hand side of Eq. (4) involves the bid, b_i, the distribution of the bids, G, and the density of the bids, g. GPV observe that if we have access to a large number of independent repetitions of the same auction, then both G and g can be consistently estimated using standard techniques. Given estimate $\hat{G}$ and $\hat{g}$ of G and g, we can form an estimate $\hat{v}_{i,t}$ of bidder i's private information $v_{i,t}$ in auction t by evaluating the empirical analogue of Eq. (4):

$$\hat{v}_{i,t} = b_{i,t} + \frac{\hat{G}(b_{i,t})}{(N-1)\hat{g}(b_{i,t})}. \quad (5)$$

To summarize, the estimator proposed by GPV is as follows:

1. Given bids $b_{i,t}$ for $i = 1,\ldots,N$ and $t = 1,\ldots,T$, estimate the distribution and density of bids $\hat{G}$ (b) and $\hat{g}$ (b).
2. Compute $\hat{v}_{i,t}$ for $i = 1,\ldots,N$ and $t = 1,\ldots,T$ using Eq. (5). Use the empirical cdf of the $\hat{v}_{i,t}$ to estimate F.

This procedure is attractive for three reasons. First, it does not impose parametric assumptions

on F during estimation. Since the economist is likely to have poor a priori information about the distribution of values, this is desirable for empirical work. Second, the procedure described above is computationally simple to implement since it does not require evaluation of ϕ and ϕ'. Finally, it is possible to demonstrate that $F(v)$ is nonparametrically identified. The intuition is quite simple. As T grows arbitrarily large, the economist will be able to estimate G and g very precisely under standard regularity conditions. Equation (4) implies that for any given bid b_i we can recover the latent valuation v_i that generates this bid (that is, $v_i = \phi(b_i)$). Since the distribution of b_i is known, it can easily be demonstrated that $F(v)$ is therefore identified.

GPV also demonstrate that the first-price auction model can be tested. Given estimates $\hat{G}$ and $\hat{g}$, define $\xi(b)$ as

$$\xi(b) = b + \frac{\hat{G}(b)}{(N-1)\hat{g}(b)}.$$

Theoretical models of bidding imply that the bid function should be increasing, that is, bidders with higher valuations should submit higher bids. Therefore, if $\hat{G}(b)/(N-1)\hat{g}(b)$ is sufficiently close to $G(b)/(N-1)g(b)$, $\xi(b)$ should be monotonically increasing if the model is correctly specified. This prediction of the theory could be rejected by the data since $\hat{G}$ and $\hat{g}$ are estimated nonparametrically and do not impose a priori that $\xi(b)$ is increasing.

Generalizations and Applications

Following GPV, a large number of authors have proposed similar estimators for other auction models. In these papers, a key step is typically to rewrite the first-order conditions in terms of the equilibrium distribution of the bids (for example G and g). Next, as in Eq. (4), the economist attempts to isolate private information on the left- hand side as a function of the bids on the right-hand side. Following GPV, the economist then nonparametrically estimates the distribution of the bids from the data and recovers the latent private information by evaluating the empirical analogue of the first-order condition.

This basic algorithm often needs to be modified for different auctions. However, attempting to follow these steps as a first pass will typically take the economist a long way towards deriving an estimator. Listed in Table 1, in alphabetical order, are some recent papers which build on the insights of GPV in other auction models.

Next, in order to illustrate how these techniques are used in practice, we briefly summarize Hortaçsu (2002) who analyses bidding in Treasury bill auctions, and Bajari and Ye (2003) who test for collusion in procurement auctions.

Auctions for Treasury Bills

Hortaçsu (2002) asks how governments should conduct auctions for Treasury bills. Treasury bill auctions are an example of a multiple unit auction since large numbers of T-bills are typically sold during a single auction. Since there are multiple units, a 'bid' in a Treasury auction is a demand curve, instead of a scalar as in the example of section "The First-Price Auction". Two commonly used mechanisms for conducting a Treasury bill auction are the uniform price auction and the discriminatory auction. In a uniform price

Auctions (Applications), Table 1 Related papers

Paper	Topic
Athey and Haile (2002)	Identification in auctions
Bajari and Ye (2003)	First-price auctions with collusion
Brendstrup and Paarsch (2003)	Dutch and first-price auctions with asymmetric bidders
Campo et al. (2002)	Auctions with risk aversion
Campo et al. (2003)	Asymmetric first-price auctions with affiliated values
Flambard and Perrigne (2006)	Asymmetric first-price auctions
Hendricks et al. (2003)	Common value auction models
Hortaçsu (2002)	Treasury auctions
Li and Perrigne (2003)	Random reserve prices
Li et al. (2002)	Affiliated private values
Pesendorfer and Jofre-Bonet (2003)	Dynamic first-price auctions

auction, the auctioneer begins by aggregating all of the individual demand curves into a market demand curve. The supply curve is vertical, with an intercept equal to the number of T-bills that the government wishes to sell. The market-clearing price is determined by the intersection of the supply and demand curve. Each bidder pays his demanded quantity at the market-clearing price, analogous to a competitive market. By contrast, in a discriminatory auction, the intersection of the supply and market demand curves determines the price for the last unit purchased. Analogous to first-degree price discrimination, bidder i pays the area under his demand curve, so that the price for the first unit purchased will be higher than for the last unit purchased.

There is no general consensus about which auction mechanism should be preferred. Since the equilibria to these auctions are quite complicated, it is difficult to characterize revenue in each auction. Each year, nearly $4 trillion dollars of securities are sold in T-bill auctions. Given the size of these markets, econometrically modelling the determination of the bids and comparing revenue from alternative auction mechanisms is an interesting public policy question.

The particular market that Hortaçsu examines is the short-term (13-week) market for T-bills in Turkey. This market is run using a discriminatory auction. Hortaçsu uses the Wilson (1979) auction of shares model as a starting point for his econometric analysis. He assumes that bidders have private values. According to surveys of bidders, 42 per cent of purchases in the auctions are to meet reserve requirements imposed by the Turkish Central bank. Thirty-seven per cent of purchases are for resale in the secondary market. Ten per cent are to fulfil customer orders and ten per cent are to fulfil collateral requirements, for investment funds administered by the bank, and for buy-and-hold purposes. Other than those shares purchased for resale, the other sources of demand are probably best modelled as private values.

Let s_i denote bidder i's private information about her willingness to pay for government debt and $v_i(q, s_i)$ denote bidder i's valuation for the qth unit. Assume that private information is distributed i.i.d. $s_i \sim F(s)$. Let $y_i(p)$ denote the demand curve submitted by bidder i. Hortaçsu assumes that $y_i(p)$ is strictly decreasing and differentiable. If there are N bidders and Q units of debt for sale, the market-clearing price p^c will satisfy

$$Q = \sum_i y_i(p^c).$$

The cdf of the market-clearing price, conditional on i's bid function $y_i(p)$ is

$$\begin{aligned} H(p, y_i(p)) &= \Pr\left\{ y_i(p) \le Q - \sum_{j \neq i} y_j(p) \right\} \\ &= \Pr\{p^c \le p | \, y_i(p)\}. \end{aligned} \tag{6}$$

Equation (6) is analogous to a residual supply curve. The term $H(p,y_i(p))$ is the probability that the market-clearing price will be less than p given i's own bid, $y_i(p)$. However, unlike a residual supply curve in a model with certainty, the bidder has to take into account her uncertainty about the bids of others.

Given a bid $y_i(p)$, the surplus that a bidder gets, conditional on p^c is equal to

$$\int_0^{y_i(p^c)} v_i(q, s_i) \mathrm{d}q - \int_0^{y_i(p^c)} y_i^{-1}(q) \mathrm{d}q.$$

There are two terms in the above sum. The first term is the integral of $v_i(q, s_i)$ from 0 to $y_i(p^c)$. This is bidder i's valuation for the units that she wins. The second term is the integral of i's inverse demand curve. This determines the total payment that i just made for the units that she won. Therefore, i's expected profit from submitting a bid of $y_i(p)$ is equal to

$$\int_0^{\infty} \left\{ \int_0^{y_i(p^c)} \left\{ v_i(q, s_i) - y_i^{-1}(q) \right\} \, \mathrm{d}q \right\} \mathrm{d}H(p^c, y_i(p)).$$

Following Wilson (1979), the first-order condition for maximization implies that

$$v_i(y_i(p), s_i) = p + \frac{H(p, y_i(p))}{\frac{\partial}{\partial p} H(p, y_i(p))}. \tag{7}$$

That is, a bidder's valuation will be equal to the price on the submitted demand curve plus a bid-shading factor, $H(p, y_i(p))/(\partial/\partial p)H(p, y_i(p))$. Just as in the first- price auction example in section "The First-Price Auction", Hortaçsu notes that $H(p, y_i(p))$ is the cdf of the equilibrium distribution of bids given $y_i(p)$. Given a large number of repetitions of the same, or similar auctions, this object can be estimated from the observed bidding data. And, similar to the first-price auction example above, an estimate of bidder i's valuation, $v_i(y_i(p), s_i)$ can be recovered by evaluating the empirical analogue of Eq. (7). While the econometric details are somewhat involved, a key economic insight was expressing the first-order conditions in terms of a function of the bids which, in principle, can be recovered from the data.

Using his estimates of bidder valuations, Hortaçsu examines two applied questions. The first is to explore the impact of reserve requirements on bidding behaviour. He constructs a variable, $\%SHORTFALL_{i,t-1}$, which is the fraction of orders in the previous Treasury auction that were unfulfilled. He finds that when bidders have a large shortfall in previous auctions, they are more likely to bid aggressively in upcoming auctions. Using his survey on bidder demands, he interprets this as derived demand from satisfying reserve requirements to hold a required portfolio of Turkish Treasury notes. For instance, he finds that the R^2 of a regression of the intercept of the submitted bid function on $\%SHORTFALL_{i,t-1}$, $\%SHORTFALL_{i,t-2}$ and an auction fixed effect is 0.61. Bidder-fixed effects only increase R^2 to 0.64.

A second applied question Hortaçsu examines is whether a uniform price auction would generate increased revenue. This is complicated to answer since changing to a uniform price auction would generate an entirely new equilibrium in this market. However, Hortaçsu demonstrates that it is possible to construct a simple upper bound on revenue given estimates of $v_i(q, s_i)$ for $i = 1,\ldots, N$. Since bidders typically engage in demand reduction in a uniform price auction, they will bid at most $v_i(q, s_i)$ so that $v_i(q, s_i)$ is an upper bound on i's bid. Assuming that this upper bound is binding for all bidders, he generates an upper bound on the market-clearing price in the auction. Using his structural estimates, Hortaçsu finds that switching to a uniform price auction would generate a revenue loss of at least 3.8 per cent on average in the auctions in his sample.

Hortaçsu therefore argues that the discriminatory price auction generates higher revenue since bidders are being forced to pay the area under their demand curves. Even after accounting for changes in the strategic incentives to shade bids, discriminatory auctions generate more revenue. However, this conclusion is subtle. Recall that bids are the steepest when shortfalls are the highest. It is hard to argue that forcing banks to hold Turkish Treasury debt is optimal for securing deposits. More likely, this policy was implemented in order to guarantee that there is a constant demand for government debt even if the government engages in irresponsible fiscal or monetary policies. These results suggest that the reserve requirements plus the discriminatory mechanism may be imposing a burden on the banking sector by forcing banks to hold more than the optimal number of domestic T-bills.

Collusion Application

Next, we briefly discuss an application by Bajari and Ye (2003) that tests for collusive bidding behaviour in procurement auctions. Bid rigging is an important antitrust problem. For instance, Pesendorfer (2000) notes that 55 per cent of the criminal antitrust cases filed by the US Department of Justice involved bid rigging. One well-known example of bid rigging was the 'concrete club' in New York where organized crime figures placed an implicit 'tax' of two per cent on every ton of concrete used in certain construction jobs in the 1980s. However, the costs of collusion were likely much larger than two per cent. Mafia informer Sammy 'The Bull' Gravano, who was involved in bid-rigging in the concrete industry, stated 'If one of them (contractor) gets a contract for, say, thirteen million, the next thing you know, after he knows he's got it, he jacks up the whole thing before it's over to a sixteen- or seventeen-million-dollar job. Now he's increased the cost 33 per cent. So our greed (the Mafia) is compounded by the greed of them so-called legitimate guys (contractors)' (Maas 1997, p. 271).

While bid-rigging is an important antitrust problem, it can be difficult to detect. Bajari and Ye (2003), expanding on the methods in section "The First-Price Auction", and on the work of Porter and Zona (1993, 1999), propose three statistical tests that can be used to potentially detect bid rigging in procurement auctions. Certainly, no test for bid-rigging can hope to be foolproof. However, it may be a basis for determining which sets of bids are most worrisome and whether further investigation of certain firms is warranted.

Bajari and Ye apply their methods to a set of contracts in the highway construction industry for 'seal coating' jobs in Minnesota, North Dakota and South Dakota. Seal coating is a type of highway repair that attempts to extend the life of the road by sealing surface cracks. The surface of the highway is initially sprayed with a coating of oil. Next, a 'chip spreader' distributes a uniform layer of sand and aggregate on the road. Finally, rollers are used to bind the oil, sand and aggregate. Bidding is conducted using sealed bids. While there are a large number of fringe firms in the industry, the market is dominated by a few large bidders that regularly compete against each other. Since all of the bids are publicly available shortly after they are submitted, collusion has occurred in seal coating in many markets. Bajari and Ye note that three of the largest bidders in their data have been fined for previous attempts to rig bids. The owner of the largest firm in the data set served prison time for a bid rigging conviction.

Bajari and Ye consider a first-price auction model similar to the example discussed in section "The First-Price Auction". However, they drop the assumption that all bidders are *ex ante* identical. In the construction industry, they argue it is important to allow for asymmetric bidders for three reasons. First, transportation costs are substantial in this market so that firms located closest to the project will tend to have lower cost. Second, there is a skewed size distribution of firms in the industry. Therefore, it is important to allow for firm specific difference in productivity. Third, project backlog increases the opportunity cost of taking on additional work and is likely therefore to be an additional source of *ex ante* asymmetries.

In the model, N firms compete for a contract to build a single and indivisible public works project. Firm i's cost to complete the project, c_i, is a random variable with cumulative distribution function $F_i(\cdot\,;z_i,\theta_i)$ and probability density function $f_i(\cdot\,;z_i,\theta_i)$. Here z_i reflects publicly observed cost shifters from firm i. For instance, in the application, these include distance to the project, a firm fixed effect to capture differences in productivity, backlog at the time bids are submitted and an engineering cost estimate. The term θ_i is a set of firm specific parameters. In the model, firm i is risk neutral and has profits of $b_i - c_i$ if it is the low bidder and zero otherwise.

Let $G_i(b;z)$ be the equilibrium distribution of bids submitted by firm i. Note that the distribution of the bids depends on $z = (z_1,\ldots,z_N)$, the publicly observed information for all firms in the industry. Then i's expected profits from submitting a bid of b_i when i's costs are c_i is equal to

$$(b_i - c_i)\prod_{j\neq 1}\big(1 - G_j(b_i;z)\big) \tag{8}$$

It can easily be shown that the first-order condition to the model must satisfy

$$c_i = b_i - \left[\sum_{j\neq 1}\frac{g_j(b_i;z)}{1-G_j(b_i;z)}\right]. \tag{9}$$

As in section "The First-Price Auction", if the economist has estimates of $\hat{G}_i$ and $\hat{g}_i$, it is possible to generate an estimate of c_i by evaluating the empirical analogue of the above equation for all bidders in the sample.

Bajari and Ye (2003) propose three tests for collusive bidding. We next describe the basic spirit of these tests, referring the interested reader to the text for complete details. The first test for competitive bidding is that conditional on z, the bids of all firms $I = 1,\ldots,N$ must be distributed independently. This is a fairly robust prediction of the theory of competitive bidding and is in fact more general than the particular model described above. Because bidders have private information which is independently distributed, their bids, which are a deterministic function of this private

information, must also be independently distributed. Obviously, one limitation of such a test is if some component of z is observed by the firms, but not by the econometrician. Following Porter and Zona (1993, 1999), their estimation strategy allows for the inclusion of an auction-specific fixed effect. Thus, they control for project specific cost shifters which are common to all of the firms.

Second, they demonstrate that the equilibrium distribution of competitive bids must be exchangeable. Let π be a permutation of the bidder identities $\{1,\ldots,N\}$, that is, a one-to-one map from $\{1,\ldots,N\}$ to $\{1,\ldots,N\}$. If the equilibrium bid function is unique, the bid distribution must be exchangeable: that is, $G_i(b; z_1; z_2; z_3, \ldots, z_N) = G_{\pi(i)}(b; z_{\pi(1)}; z_{\pi(2)}; z_{\pi(3)}; \ldots, z_{\pi(N)})$. In words, exchangeability means that if you permute the cost shifters of all the bidders, then the equilibrium bids must also permute in a symmetric fashion. Conditional independence and exchangeability are necessary for equilibrium bidding. If other regularity conditions hold, conditional independence and exchangeability are also sufficient for competitive bidding: that is, the economist can reverse engineer a competitive bidding model that rationalizes the observed bids.

Porter and Zona (1993, 1999) study the bidding behaviour of known cartels in construction and in the supply of school milk. Many of the irregular patterns of bidding that they describe can be characterized as failures of conditional independence and exchangeability. For instance, the bids of cartel members are more correlated with each other than with non-cartel members. Also, cartel members do not shift their bids aggressively in response to shifts in the z_i of other cartel members which is a failure of exchangeability.

Bajari and Ye (2003) test for conditional independence and exchangeability in their data set. Given the limited number of observations available to them, they test these conditions in a regression framework. Essentially, they run a regression of b_i on z_i and z_{-i}, including auction fixed effects and bidder fixed effects. Conditional independence is tested by asking whether the fitted residuals from bidder i's bid function is correlated with the fitted residuals from $j \neq i$'s bid function. Exchangeability is formulated as a test of the equality of certain regression coefficients. In total, 46 separate hypothesis tests are conducted. Forty-one of these tests are consistent with the implications of competitive bidding (that is, conditional independence and exchangeability). Therefore, they argue that most of the bids in the market appear to be competitive. However, reduced form tests suggest that bidding by two coalitions of firms appear to be suspicious. They label these coalitions 'candidate cartels'. Interestingly, all of the members of the candidate cartels had previously been convicted of bid rigging.

The third and final test for bid rigging uses structural estimates based on Eq. (9). Bajari and Ye consider a non-nested hypothesis test between three models. Model M1 is that the data-generating process is the no collusion model. Model M2 is that the first candidate cartel is engaged in efficient collusion, but that other firms in the industry are competitive. Model M3 is that the second candidate cartel engages in bid rigging. The costs c_i can be estimated under each of these three alternatives using the empirical analogue of Eq. (9). The different models generate different first-order conditions and hence, different estimated costs, c_i.

Bajari and Ye then ask which set of markups is 'most reasonable'. To answer this question, they consulted with two managers at one of the biggest firms in this market (which was not in a candidate cartel). From each manager, they elicited their beliefs about the distribution of markups in this industry. Bajari and Ye argue that it is reasonable to suppose that these managers have informative priors about markups for two reasons. First, all bidders in this industry must be bonded. The bonding companies are contractually liable to complete the project if the contractors go bankrupt. Contractors are typically required to give weekly profit and loss statements to the bonding companies. The bonding companies are therefore well informed about profit margins for firms in the industry. Profit margins in the industry are a common topic of conversation between contractors and bonding companies and are one source of information.

Second, the contractors in this industry compete against each other quite frequently and over many years. The contractors have access to similar cost information and study the bids of competing contractors in detail after the bids are publicly opened. Given that contractors closely follow cost conditions and bids in the industry, they will have a lot of information about their competitors' markups. There is an issue, of course, about whether the contractors would lie about their beliefs. However, Bajari and Ye shared their estimates with the contractor, which included empirical analysis of the behaviour of competing firms. Lying about the industry would reduce the value of these estimates. Also, the information from the contractor that was verifiable from external sources about the industry did seem to be accurately reported.

The stated beliefs of the experts were quite close. Below, we average the elicited beliefs from the contractors:

$$\begin{aligned} &25\text{th percentile} = 3\%50\text{th percentile} \\ &\quad = 5\%75\text{th percentile} \\ &\quad = 7\%99\text{th percentile} = 15\%. \end{aligned} \tag{10}$$

For example, the 25th percentile of the bids has a markup of three per cent and the median bid has a markup of five per cent.

Table 2 shows the estimated distribution of markups from the three alternative structural models, M1, M2 and M3.

Bajari and Ye note that the markups under M1 (competitive bidding) correspond most closely to the elicited prior beliefs. The markups under models M2 and M3 seem to be too large, particularly on the tails. They argue that this is evidence against the collusive models since they generate markups that seem implausibly large compared to the beliefs of an informed party. Bajari and Ye formalize this intuition by posing the selection of M1, M2 or M3 as a problem in statistical decision theory. As the table above suggests, the competitive model M1 is most favoured. Therefore, they cautiously interpret the data as being consistent overall with non-collusive behaviour.

Auctions (Applications), Table 2 Distribution of markups under alternative models

Percentile	M1	M2	M3
10	0.01229	0.01273	0.0114
20	0.01597	0.01818	0.0182
30	0.02077	0.02422	0.0256
40	0.02536	0.03201	0.0343
50	0.03329	0.04126	0.0447
60	0.04227	0.05434	0.0584
70	0.05692	0.0754	0.0930
80	0.1000	0.1621	0.1756
90	0.2381	0.3354	0.5826

Conclusion

In this short survey, I have attempted to provide an overview of recent empirical papers concerning auctions. Many recent papers build on the pioneering work of Guerre, Perrigne and Vuong (2000). A key insight of this paper was that a first-price sealed-bid auction model can be simply estimated using a two-step procedure. In the first step, the economist flexibly estimates the empirical distribution of the bids. In the second step, the economist evaluates the empirical analogue of the first-order condition for utility maximization. The method of Guerre, Perrigne and Vuong estimates the structural primitives of the model without imposing ad hoc parametric restrictions. We also discussed two applications of these recently developed estimators. Hortaçsu (2002) studied bidding in Treasury auctions in Turkey. His model predicted that discriminatory auctions generate higher revenue than uniform price auctions. Bajari and Ye (2003) applied these methods to test for collusion in sealed-bid auctions. They applied these methods to searching for suspicious bidding patterns in a market where the largest firms had recently been sanctioned for collusion.

See Also

- ▶ Auctions (Empirics)
- ▶ Auctions (Experiments)
- ▶ Auctions (Theory)
- ▶ Cartels

▶ Epistemic Game Theory: Incomplete Information
▶ Game Theory
▶ Non-parametric Structural Models

Bibliography

Athey, S., and P. Haile. 2002. Identification of standard auction models. *Econometrica* 70: 2107–2140.
Athey, S., and P. Haile. 2007. Nonparametric approaches to auctions. In *Handbook of econometrics*, ed. J.J. Heckman and E.E. Leamer, vol. 6. Amsterdam: North- Holland.
Bajari, P., and A. Hortaçsu. 2003. The winner's curse, reserve prices, and endogenous entry: Empirical insights from eBay auctions. *RAND Journal of Economics* 34: 329–355.
Bajari, P., and L. Ye. 2003. Deciding between competition and collusion. *Review of Economics and Statistics* 85: 971–989.
Brendstrup, B., and H. Paarsch. 2003. Nonparametric estimation of Dutch and first- price, sealed-bid auction models with asymmetric bidders. Working paper, University of Iowa.
Campo, S., E. Guerre, I. Perrigne, and Q. Vuong. 2002. Semiparametric estimation of first-price auctions with risk averse bidders. Working paper, Pennsylvania State University.
Campo, S., I. Perrigne, and Q. Vuong. 2003. Asymmetry in first-price auctions with affiliated private values. *Journal of Applied Econometrics* 18: 197–207.
Donald, S., and H. Paarsch. 1993. Piecewise pseudo-maximum likelihood estimation in empirical models of auctions. *International Economic Review* 34: 121–148.
Flambard, V., and I. Perrigne. 2006. Asymmetry in procurement auctions: Evidence from snow removal contracts. *Economic Journal* 116: 1014–1036.
Guerre, E., I. Perrigne, and Q. Vuong. 2000. Optimal nonparametric estimation of first-price auctions. *Econometrica* 68: 525–574.
Hendricks, K., and R. Porter. 2007. Lectures on auctions: An empirical perspective. In *Handbook of industrial organization*, ed. M. Armstrong and R. Porter, vol. 3. Amsterdam: Elsevier.
Hendricks, K., J. Pinkse, and R. Porter. 2003. Empirical implications of equilibrium bidding in first-price, symmetric, common value auctions. *Review of Economic Studies* 70: 115–145.
Hirano, K., and J. Porter. 2003. Asymptotic efficiency in parametric structural models with parameter dependent support. *Econometrica* 71: 1307–1338.
Hong, H., and H.J. Paarsch. 2006. *An introduction to the structural econometrics of auction data*. Cambridge, MA: MIT Press.
Hong, H., and M. Shum. 2002. Increasing competition and the winner's curse: Evidence from procurement. *Review of Economic Studies* 69: 871–898.
Hortaçsu, A. 2002. Mechanism choice and strategic bidding in divisible good auctions: An empirical analysis of the Turkish treasury auction market. Working paper, University of Chicago.
Li, T., and I. Perrigne. 2003. Timber sale auctions with a random reserve price. *Review of Economics and Statistics* 85: 189–200.
Li, T., I. Perrigne, and Q. Vuong. 2002. Structural estimation of the affiliated private value auction model. *RAND Journal of Economics* 33: 171–193.
Maas, P. 1997. *Underboss: Sammy the Bull Gravano's story of life in the Mafia*. New York: HarperCollins.
Paarsch, H. 1992. Deciding between common and private values paradigms in empirical models of auctions. *Journal of Econometrics* 51: 191–215.
Pesendorfer, M. 2000. A study of collusion in first-price auctions. *Review of Economic Studies* 67: 381–411.
Pesendorfer, M., and M. Jofre-Bonet. 2003. Estimation of a dynamic auction game. *Econometrica* 71: 1443–1489.
Porter, R., and J.D. Zona. 1993. Detection of bid rigging in procurement auctions. *Journal of Political Economy* 101: 518–538.
Porter, R., and J.D. Zona. 1999. Ohio school milk markets: An analysis of bidding. *RAND Journal of Economics* 30: 263–288.
Wilson, R. 1979. Auctions of shares. *Quarterly Journal of Economics* 93: 675–689.

Auctions (Empirics)

Isabelle Perrigne and Quang Vuong

Abstract

The structural analysis of auction data relying on game theoretic models has undergone a tremendous development since the mid-1990s. This article reviews some important contributions for first-price and ascending auctions. It stresses identification of the structure and the development of tractable econometric methods, while addressing bidders' asymmetry, common value, bidders' risk aversion, endogenous entry, dynamic and multi-unit auctions as well as the choice of the reserve price and the auction mechanism. Various domains are studied, such as auctions of timber, gas lease, treasury bills, agricultural products, electricity and construction procurements.

Keywords

Affiliated private value model; Asymmetric information; Auctions; Bayesian Nash equilibrium; Bidding; Collusion; Game theory; Log-normal distribution; Maximum likelihood; Maximum likelihood; Nonlinear least squares; Nonparametric estimation; Nonparametric methods; Reserve price; Risk aversion

JEL Classification

D44

Auctions and procurements are widely used market mechanisms for allocating public contracts, financial securities, agricultural products, natural resources, artwork, and electricity, to name a few commodities. Recent years have also witnessed the developments of auction websites and business-to-business auctions. In general, auctions have well-defined rules that can be captured by an economic model. Relying on the concept of the Bayesian Nash equilibrium, game theory has greatly contributed to the modelling of auctions, where a seller or buyer faces a limited number of bidders who behave strategically. The auction is typically an incomplete information game where the asymmetry of information between the seller/buyer and the bidders and among the bidders themselves plays a crucial role.

While auctions are largely used in economic life and data are rich and accessible, until recently the empirical analysis of auction data has been confined to testing some predictions generated by game theoretic models. One influential example of the reduced form approach is the work by Porter and his coauthors on the role of private information in oil and gas auctions, as surveyed in Porter (1995). This approach has also been used to test for collusive behaviour in timber and milk auctions. Although important, this approach does not allow for policy evaluations that require knowledge of the informational structure of the game such as the choice of the reserve price and the auction mechanism that would generate greater revenue for the seller/buyer.

The structural approach addresses such questions by assuming that observed bids are the equilibrium bids of some auction model. Specifically, $b_i = s_i(v_i)$ where b_i and v_i are bidder's i (observed) bid and (unobserved) private information, respectively, and $s_i(\cdot)$ is bidder's i equilibrium strategy in the corresponding auction game. Bidders' private information is assumed to be derived from some distribution that is common knowledge to all bidders. This distribution and the bidders' preferences are the key elements that explain bidding behaviour. They are the structural elements of the induced econometric model for the observed bids. The structural approach then exploits the equilibrium relations $b_i = s_i(v_i)$ to recover bidders' private information, which can be exploited for policy purposes. A major difficulty in implementing this approach arises from the numerical complexity or the implicit form of the equilibrium strategies. Of its nature, the structural approach raises challenging questions. One question is related to identification, namely, whether the auction structure can be uniquely recovered from observables while minimizing parametric restrictions. This question relates to whether auction models can be distinguished from observables. A second question concerns the model validity, namely, whether an auction model imposes testable restrictions on observables. A third difficulty is to develop tractable estimation methods. Since ascending (English) auctions and first-price sealed-bid auctions involve different equilibrium strategies and different identification and estimation problems, they are treated separately.

Econometrics of First-Price Auctions and Applications

Two kinds of methods can be distinguished. Direct methods start from a parameterization of the private information distribution $F(\cdot)$ and sometimes require the computation of equilibrium strategies. Indirect methods exploit the first-order condition(s) to estimate $F(\cdot)$ from the observed bid distribution without computing the equilibrium strategies. Direct methods require explicit forms for the equilibrium strategies, while indirect methods can be considered when no explicit form exists. The structural approach was initiated

by Paarsch (1992) using a direct method to analyse tree planting contract auctions with symmetric bidders. If the latent distribution is parameterized as $F(\cdot;\ \theta)$, then $b_i = s(v_i;\ \theta)$, which is distributed as $G(\cdot;\ \theta) = F[s^{-1}(\cdot;\ \theta);\ \theta]$. This raises two difficulties. First, a limited number of distributions lead to tractable equilibrium strategies. Second, the standard regularity conditions of maximum likelihood (ML) estimation are violated because the bid distribution support depends on θ. Paarsch and coauthors have extended ML estimation to this problem. Laffont et al. (1995) propose an alternative direct method based on simulations while analysing Dutch auctions of vegetables. This method allows a large family of distributions to be entertained. It exploits the revenue equivalence theorem for independent private value models to write the expectation $m_\ell(\theta)$ of the winning bid in a Dutch auction as the expectation of the second highest value. The authors develop a simulated nonlinear least squares estimator based on minimizing $Q_L(\theta) = (1/L)\Sigma_{\ell=1}^{L}\left[b_\ell^w - m_\ell(\theta)\right]$, where $m_\ell(\theta)$ is replaced by a simulator, L is the number of auctions and b^w is the winning bid, while correcting for its inconsistency. This idea has been extended by others when the expected winning bid can be simulated. This limits the number of models to be considered. Bayesian estimation methods, though computationally demanding, have also been developed.

In contrast, the indirect method initiated by Guerre et al. (2000) requires neither the computation nor the simulation of equilibrium strategies. It uses the differential equation (s) or first-order condition(s) to express each private value as a function of its corresponding bid. Within the symmetric independent private value paradigm, the differential equation is $s'(v_i) = [v_i - s(v_i)](I-1)[f(v_i)/F(v_i)]$, where I is the (known) number of bidders, $s'(\cdot)$ is the derivative of $s(\cdot)$ and $f(\cdot)$ is the private value density. Because $b_i = s(v_i)$, bids are also i.i.d. with $G(b) = F[s^{-1}(b)] = F(v)$ leading to $g(b) = f(v)/s'(v)$. Hence, the differential equation can be written as

$$v_i = b_i + \frac{1}{I-1}\frac{G(b_i)}{g(b_i)} \equiv \xi(b_i, G, I). \qquad (1)$$

Relying on (1), the authors show that the model is nonparametrically identified: that is, one can recover uniquely the distribution $F(\cdot)$ from the observed bid distribution without parametric restrictions. Moreover, they derive the restrictions imposed by the model on observables: that is, bids must be i.i.d. (since private values are i.i.d.) and $\xi(\cdot)$ should be strictly increasing (since $s(\cdot)$ is strictly increasing). These two restrictions can be used to test the validity of the model. Equation (1) calls for a two-step estimation procedure. The first step consists in estimating nonparametrically $G(\cdot)$ and $g(\cdot)$, while the second step estimates nonparametrically $f(\cdot)$ from the estimated private values $\widehat{v}_i$ using (1). In practice, auctioned goods are heterogeneous. Observed characteristics can be introduced in the econometric model by writing (1) with conditional bid distribution and density. Nonparametric estimation can be a drawback when a limited number of auctions is available and/or when the number of exogenous variables is relatively large. It can, however, provide a preliminary estimate of the underlying density, which can be used later to specify $F(\cdot)$ when using a parametric two-step estimation procedure.

In addition to not parameterizing $F(\cdot)$, the indirect method does not require an explicit form for the equilibrium strategy, as it relies on the first-order condition(s). The method provides key insights on questions at the core of the structural approach, as discussed above. It can be easily extended to the case of a binding reserve price, where the number of actual (observed) bidders is smaller than the number I of potential bidders as only bidders with private values above the reserve price effectively participate. Alternatively, the seller may not announce his reserve price, keeping it secret as in timber and wine auctions. Although the equilibrium strategy in such a model does not have an explicit form, the above method allows a simple expression to be obtained for the inverse equilibrium strategy, which can be used to develop a two-step estimation procedure as above. Likewise, the method can be easily extended to situations in which only the winning bids are observed, as in Dutch auctions, which are widely used for agricultural products such as vegetables and flowers.

Independence among private values can be restrictive. One can expect some affiliation or positive correlation among private values and some common value v affecting all bidders' utilities, that is, bidder's i utility becomes $v_i = U(\sigma_i, v)$. In the private value paradigm $v_i = \sigma_i$, while in the pure common value paradigm $v_i = v$. The vector $(\sigma_1, \ldots, \sigma_I, v)$ is distributed as $F(\cdot, \ldots, \cdot)$, which is affiliated and exchangeable in its first I arguments under bidders' symmetry. Affiliation means that, if one bidder values the auctioned object highly, other bidders are also likely to value it highly. In the common-value model bidders receive signals about the value of the object, which is unknown at the time of the auction. This model has been widely used to explain bidding behaviour in gas lease auctions where firms have imperfect information about the amount of oil. The general framework is considered by Laffont and Vuong (1996), who study the problem of identification and theoretical restrictions. They show that any symmetric affiliated value model is observationally equivalent to some symmetric affiliated private value (APV) model because $U(\cdot)$ is unidentified, as any dependence across utilities arising from v can be replaced by a dependence among private values. Similarly, the pure common value is unidentified from observed bids. If some additional information is available, such as the *ex post* common value, identification can be achieved. On the other hand, the symmetric APV model is identified.

Regarding estimation, a two-step estimation procedure can be developed. Let $B_1 = s(y_1)$ with $y_1 = max_{j \neq 1} \sigma_j$. When $v_i = \sigma_i$, (1) becomes

$$v_i = b_i + \frac{G_{B1|b_1}(b_i|b_i)}{g_{B1|b_1}(b_i|b_i)} \equiv \xi(b_i, G). \qquad (2)$$

Regarding theoretical restrictions, $\xi(\cdot)$ needs to be strictly increasing and the bid distribution $G(\cdot, \ldots, \cdot)$ must be affiliated and exchangeable. An interpretation of the APV model is that affiliation arises from some latent variable v. Building on this interpretation, Li et al. (2000) propose a model with private information conditionally independent upon some common component. Specifically, each piece of private information is the product of two unobserved independent components, one specific to the auctioned object and common to all bidders, the other specific to each bidder, that is, $\sigma_i = v\eta_i$. Hence, $\log \sigma_i = \log x + \log \varepsilon_i$ with $\log x = [\log v + \mathrm{E}(\log \eta)]$ and $\log \varepsilon_i = [\log \eta_i - \mathrm{E}(\log \eta)]$ showing that $\log \varepsilon_i$ can be interpreted as an error term in a measurement error model with $\log x$ unobserved. Because the v_i can be recovered from (2) when $v_i = \sigma_i$, the densities for $\log x$ and $\log \varepsilon$ are nonparametrically identified and estimated with the use of characteristic functions. When $v_i = v$, (2) gives $\mathrm{E}[v \mathrm{j} \sigma_1 = \sigma, y_1 = \sigma]$. Under loglinearity of the latter, that is, $\log \mathrm{E}[v|\sigma_1 = \sigma, y_1 = \sigma] = C + D \log \sigma$, the pure common value model is identified up to location and scale. It is important to test whether a common value or private value paradigm is the more appropriate. Recent developments exploit how $\mathrm{E}[v|\sigma_1 = \sigma; y_1 = \sigma]$ varies with the number of bidders to formulate such tests.

Several auction data provide evidence of bidders' asymmetry, which can arise from, for example, different firms' sizes, different access to information such as the drainage auctions, and different capacity constraints and locations as in construction procurements. Collusion may also lead to asymmetry as a cartel of bidders behaves differently from other bidders. Asymmetry is *ex ante* known to all bidders. A common feature of asymmetric auction models is that they lead to intractable systems of differential equations. Hence, the direct approach is difficult to implement as it requires the numerical determination of the equilibrium strategies for any trial parameter value. Let $F_1(\cdot), \ldots, F_I(\cdot)$ be the private value distributions of the I bidders whose identities are observed. For simplification, independent private values are considered, though the method can be easily extended to affiliated private values. Let $G_1(\cdot), \ldots, G_I(\cdot)$ be the corresponding bid distributions. The intractable system of differential equations can be rewritten as

$$v_i = b_i + \frac{1}{\sum_{j \neq i} \frac{g_j(b_i)}{G_j(b_i)}}, i = 1, \ldots, I. \qquad (3)$$

This method has been used to analyse joint bidding in gas lease auctions and snow removal procurements, where asymmetry arises from a firm's location relative to contract location.

Bidders' risk neutrality is often assumed because the value of the object is small relative to bidders' assets. Recent studies have suggested that bidders may be risk averse in timber auctions. The experimental literature has noted a tendency to bid above the Bayesian Nash equilibrium, which can be rationalized by risk aversion. In a private value framework, the bidder's utility becomes $U(v_i - b_i)$ with $U(\cdot)$ strictly increasing and concave. Campo et al. (2006) study the identification and estimation of risk aversion. Using an indirect approach and omitting wealth to simplify, the differential equation defining the equilibrium strategy becomes

$$v_i = b_i + \lambda^{-1}\left(\frac{1}{I-1}\frac{G(b_i)}{g(b_i)}\right) \equiv \xi(b_i, U, G, I), \quad (4)$$

where $\lambda^{-1}(\cdot)$ denotes the inverse of $\lambda(\cdot) = U(\cdot)/U'(\cdot)$. The model is not identified only from observed bids. In fact, any bid distribution can be rationalized by a constant relative or absolute risk aversion model. Additional restrictions, such as parameterizing either the utility function or the private value distribution, are not sufficient to identify the model as an increase in the risk aversion parameter can be compensated by a shrinkage of all the quantiles of $F(\cdot)$. Consequently, the authors parameterize a single quantile of $F(\cdot)$ to achieve identification of the model while exploiting auction heterogeneity. Under parameterization of $U(\cdot)$ and a conditional quantile, (4) at any quantile provides an estimating equation for the parameters of the utility function and the quantile of $F(\cdot)$. The method can be easily extended to affiliated private values and bidders' asymmetry in private values. Alternatively, if the number of bidders is exogenous, that is, $F(\cdot)$ is independent of I, nonparametric identification can be achieved. More generally, exclusion restrictions help in identifying the model. Regarding asymmetry, bidders may have heterogeneous preferences, that is, they may have different attitudes towards risk given their assets, experience, and so on. Thus, (4) evaluated at any quantile for two different bidders provides additional identifying restrictions since the corresponding quantile of $F(\cdot)$ is equal. Construction procurement data show that firms with more experience tend to be less risk averse. Risk aversion has important implications for several policy issues including the announcement of the reserve price and the auction format. These results allow more advanced auction models to be considered, in which risk aversion plays a key role. Examples includes stochastic values when uncertainties affect bidders' *ex post* value and financially constrained bidders.

Identical commodities such as treasury bills and electricity are sold sometimes through multiunit auctions. A bidder acquires a share of the quantity supplied. Each bidder submits several (quantity, price) pairs. Hortaçsu (2002) studies discriminatory share auctions of treasury bills while considering private values in light of empirical evidence. Each bidder strategy is a demand function $y(p, \sigma_i)$ where σ_i is bidder's i private information. The clearing price P_c equates the bidder's demand function with the residual supply curve $Q - \sum_{j\neq i}^{I} y(p, \sigma_j)$, where Q is the total supply. Let $G(p, x)$ be the distribution of the residual supply faced by bidder i at price p given $y(p, \sigma_i) = x$, that is,

$$G(p,x) = \mathrm{pr}\left[x \le Q - \sum_{j\neq i}^{I} y(p,\sigma_j) \,|\, y(p,\sigma_i = x)\right] = \mathrm{pr}[P_c \le p | y(p,\sigma_i) = x].$$

The optimal bid p for the quantity $y(p; \sigma_i)$ is

$$v[y(p,\sigma_i),\sigma_i] = p + \frac{G[p, y(p,\sigma_i)]}{\partial G[p, y(p,\sigma_i)]/\partial p},$$

where $v[y(p, \sigma_i), \sigma_i]$ is bidder's i marginal utility from winning the $y(p, \sigma_i)$th unit. With the use of a re-sampling strategy to estimate $G(\cdot, \cdot)$, the results are used to compare the discriminatory price

mechanism with the uniform price mechanism. The problems of identification of the private information distribution and the restrictions imposed by the model on observables remain to be solved. This method has also been applied to electricity auctions.

The preceding developments ignore dynamic considerations, while bidders frequently participate in several auctions over time. Jofre-Bonet and Pesendorfer (2003) consider a dynamic auction to analyse highway construction procurements where previously won uncompleted contracts introduce capacity constraints affecting firms' actual costs. This involves inter-temporal optimization, while introducing asymmetry among bidders arising from different capacity constraints, location and size. If we use an indirect approach, the inverse equilibrium strategies solve

$$c_i = b_i - \frac{1}{\sum_{j\neq i} \frac{g_j(b_i)}{1-G_j(b_i)}} + \beta \sum_{j\neq i} \frac{\frac{g_j(b_i)}{1-G_j(b_i)}}{\sum_{k\neq j} \frac{g_k(b_i)}{1-G_k(b_i)}} [V_i(\omega(i)) - V_i(\omega(j))], \quad i = 1, \ldots, I, \tag{5}$$

where β is a discount factor, $\omega(i)$ is a transition function indicating the sizes and remaining times of all current projects for bidder i, $V_i(\cdot)$ is the value function determining the discounted sum of expected future profits. The system (5) is similar to (3) with cost c_i and $1 - G_j(\cdot)$ as the firm with the lowest bid wins the procurement. Because the value function can be written as a function of the bid distributions, identification comes down to whether the cost distributions and the discount factor can be uniquely recovered from observed bids. Identification is obtained when the discount factor is known. Relying on standard numerical methods to approximate the value function, a two-step parametric procedure allows us to estimate the cost distributions $F_1(\cdot), \ldots, F_I(\cdot)$.

Econometrics of Ascending Auctions and Applications

In the private value paradigm, a dominant strategy for every bidder is to exit the auction at his valuation. The bidding process ends when a single bidder remains. In the button auction model, the winning bid can be interpreted as the second highest among I values. Athey and Haile (2002) study identification of ascending auctions while emphasizing data requirements. When private values are independent and the number of bidders is observed, the transaction price is the $(I-1)$th order statistic $v^{(I-1:I)}$ whose distribution is

$$F^{(I-1:I)}(v) = \frac{I!}{(I-2)!} \int_0^{F(v)} t^{I-2}(1-t)dt,$$

from which the distribution $F(v)$ is recovered. When bidders are asymmetric and bidders' identities are known, a similar argument can be used to show that $F_1(\cdot), \ldots, F_I(\cdot)$ are identified. Nonparametric estimation can be performed. The problem becomes complicated when one considers more general frameworks. When private values are affiliated, the winning bid is not sufficient to recover affiliation among bids and hence $F(\cdot, \ldots, \cdot)$. Additional observations are needed.

However, many ascending auctions do not match the button auction model. In practice, bidders do not continuously indicate whether they are still participating. Moreover, because bid increments are often used, bidders may fail to reveal their willingness to pay or even to bid. In the empirical literature it is agreed that, at most, the winning bid can be rationalized by the ascending auction model. An alternative approach is proposed by Haile and Tamer (2003), who formulate an incomplete model based on two simple assumptions: (*a*) bidders do not bid more than they are willing to pay; and (*b*) bidders do not allow an opponent to win at a price they can beat. These assumptions do not allow us to identify the private value distribution but provide some bounds on this distribution. Assumption (*a*) implies $b^{(i:I)} \leq v^{(i:I)}$ or equivalently $F^{(i:I)}(v) \leq G^{(i:I)}(v)$ for $i = 1, \ldots, I$. This inequality is used to construct the upper bound for $F(\cdot)$ as

$$F^U(v) = \min_{i,I} \varphi\left[G^{(i,I)}(v); i, I\right],$$

where φ(·) is a strictly increasing function defined as $F(v) = \varphi\left[F^{(i:I)}(v); i, I\right]$. Assumption (*b*) implies that all losing bidders have valuations no higher than the winning bid plus a bid increment Δ, i.e. $v_i \leq b^{(I:I)} + \Delta$ if $b_i < b^{(I:I)}$. Let $G_\Delta^{(I:I)}(\cdot)$ be the distribution of $b^{(I:I)} + \Delta$. Thus $G_\Delta^{(I:I)}(v) \leq F^{(I-1:I)}(v)$, which is used to construct the lower bound for $F(\cdot)$ as

$$F^L(v) = \max_I \varphi\left[G_\Delta^{(I:I)}(v); I-1, I\right].$$

Nonparametric estimation of $F^U(\cdot)$ and $F^L(\cdot)$ is proposed. Tight estimated bounds suggest that the data do not deviate much from the button auction model. Bounds for the optimal reserve price can also be derived. The method is illustrated on timber auction data, and can be extended to affiliated private values and asymmetric bidders.

In a common value paradigm, bidding takes a more complex form as bidders obtain information during the auction when their rivals drop out. The auction can be modelled as a game with several rounds with $I-1$ rounds indexed by $k = 0, 1, \ldots, I-2$. Bidders are indexed in the inverse order of their dropping out. Each bidder observes a signal σ_i of his value v_i. An interesting feature of the ascending common value auction is that bidder's j dropping out is useful to bidder i for evaluating his own v_i. In this game, every bidder has $I-1$ bidding functions $s_{ik}(\cdot)$; $k = 0, \ldots, I-2$. With asymmetric bidders, the equilibrium bid functions at round k are given by

$$s_{ik}(\sigma_i) = \mathrm{E}\left[v_i | \sigma_i; \sigma_j = s_{jk}^{-1}(s_{ik}(\sigma_i)),\ j = 1, \ldots, I-k,\ j \neq i, \Omega_k\right],\quad i = 1, \ldots, I-k,$$

where $\Omega_k = \{\sigma_j = s_{j,I-j}^{-1}(P_{I-j}),\ j = I-k+1, \ldots, I\}$ is the public information set containing the observed signals of the bidders who have dropped out prior to round k and P_k is the (observed dropping out) price. Thus, at round k the $I-k$ inverse bidding strategies are solutions of the system of nonlinear equations

$$P_k = E\left[v_i | \sigma_i = s_{ik}^{-1}(P_k); \sigma_j = s_{jk}^{-1}(P_k),\ j = 1, \ldots, I-k,\quad j \neq i, \Omega_k\right]. \tag{6}$$

Using log-normal distributions and a multiplicative form for v_i and σ_i, Hong and Shum (2003) develop a tractable econometric model based on (6) that is estimated by either maximum likelihood or simulated nonlinear least squares. An illustration of the method is proposed on spectrum auctions which are organized in multiple rounds.

The recent development of auction websites provides new data opportunities. Bajari and Hortaçsu (2003) analyse coin auctions within a common value framework in light of resale opportunities, while bidders face an entry cost leading to endogenous entry. Another interesting characteristic is that the reserve price can be either posted or secret. As is well known, bidding activity is concentrated at the very end of the auction. The authors show that this practice, known as 'sniping', can be explained by a two-stage game in which no bidding is an equilibrium in the first stage, while second stage bids are the equilibrium bids in a sealed-bid second-price auction. Empirical results show that bidders' entry increases with a secret reserve price.

Concluding Remarks

The structural approach to analysing bidding data has been a field of extremely active research in the recent years. It has also contributed to the development of new econometric techniques. Many interesting problems remain to be addressed. Since auction models can be viewed as simple forms of asymmetric information, one can expect that more progress will be made in the analysis of complex asymmetric information models such as contracts.

See Also

- ▶ Auctions (Applications)
- ▶ Auctions (Theory)
- ▶ Non-parametric Structural Models

Bibliography

Athey, S., and P. Haile. 2002. Identification of standard auction models. *Econometrica* 70: 2107–2140.

Bajari, P., and A. Hortaçsu. 2003. The winner's curse, reserve prices and endogenous entry: Empirical insights from eBay auctions. *Rand Journal of Economics* 34: 329–355.

Campo, S., E. Guerre, I. Perrigne, and Q. Vuong. 2006. *Semiparametric estimation of first-price auctions with risk averse bidders*. Working paper. University Park: Pennsylvania State University.

Guerre, E., I. Perrigne, and Q. Vuong. 2000. Optimal nonparametric estimation of first-price auctions. *Econometrica* 68: 525–574.

Haile, P., and E. Tamer. 2003. Inference with an incomplete model of English auctions. *Journal of Political Economy* 111: 1–51.

Hong, H., and M. Shum. 2003. Econometric models of asymmetric ascending auctions. *Journal of Econometrics* 112: 327–358.

Hortaçsu, A. 2002. *Mechanism choice and strategic bidding in divisible good auctions: An empirical analysis of the Turkish treasury auction market*. Working paper. Chicago: University of Chicago.

Jofre-Bonet, M., and M. Pesendorfer. 2003. Estimation of a dynamic auction game. *Econometrica* 71: 1443–1489.

Laffont, J.-J., and Q. Vuong. 1996. Structural analysis of auction data. *American Economic Review: Papers and Proceedings* 86: 414–420.

Laffont, J.-J., H. Ossard, and Q. Vuong. 1995. Econometrics of first-price auctions. *Econometrica* 63: 953–980.

Li, T., I. Perrigne, and Q. Vuong. 2000. Conditionally independent private information in OCS wildcat auctions. *Journal of Econometrics* 98: 129–161.

Paarsch, H. 1992. Deciding between the common and private value paradigms in empirical models of auctions. *Journal of Econometrics* 51: 191–215.

Porter, R. 1995. The role of information in US offshore oil and gas lease auctions. *Econometrica* 63: 1–27.

Auctions (Experiments)

John H. Kagel and Dan Levin

Abstract

Experiments permit rigorous investigations of auction theory generating a dialogue with theorists and policymakers. In single-unit private value auctions the revenue equivalence theorem fails, but the comparative static predictions of Nash bidding theory hold, indicating that bidders are responsive to the primary economic forces present in the theory. In single-unit common value auctions inexperienced bidders invariably suffer from a 'winner's curse', and the comparative static predictions of the theory fail, but more experienced bidders do substantially better. Recent research dealing with Internet auctions, mixed private and common value auctions and multiunit demand auctions are surveyed as well.

Keywords

Adverse selection; Auctions; Auctions (experiments); Becker–DeGroot–Marshak procedure; Common value auctions; Dutch auctions; English auctions; English clock auctions; First-price auctions; Independent private values model; Internet auctions; Learning direction theory; Mechanism design; Multiunit demand auctions; Revenue equivalence theorem; Risk aversion; Risk-neutral Nash equilibrium; Second-price auctions; Vickrey auction; Winner's curse

JEL Classification

C9

Experimental work in auctions interacts with theory, providing a basis for testing and modifying theoretical developments. It has advantages and disadvantages relative to empirical work with field data, so that we view the two as complimentary. Experimental work is used increasingly as a test bed for new auction formats such as the Federal Communication Commission's (FCC) sale of spectrum (air-wave) rights.

Until recently most of theoretical and experimental work was devoted to single-unit demand auctions. With the success of the FCC's spectrum auctions, much of the interest has shifted to auctions in which individual bidders demand multiple units. Experimental work in this area is still in its infancy. In keeping with the historical development of the field, we first report on single-unit demand auctions and then move to multi-unit demand auctions and Internet auctions.

Single-Unit, Private-Value Auctions

Initial experimental research on auctions focused on the independent private values (IPV) model investigating the revenue equivalence theorem. In the IPV model each bidder knows his valuation of the item with certainty, bidders' valuations are drawn identically and independently from each other, and bidders know the distribution from which their rivals' values are drawn (but not their values) and the number of bidders. Under the revenue equivalence theorem the four main auction formats – first- and second-price sealed-bid auctions, English and Dutch auctions – yield the same average revenue for risk neutral bidders. Further, first-price sealed-bid and Dutch auctions are theoretically isomorphic – they yield the same revenue for each auction trial regardless of risk preferences – as are second-price sealed-bid and English clock auctions. These isomorphisms are particularly attractive as it is hard to control bidders' risk preferences. These theoretical results are also quite surprising and counter-intuitive as the Dutch auction starts with a high price which is lowered until a bidder accepts at that price. And in the English auctions the price starts low and increases until only one bidder is left standing and pays the price where the next-to-last bidder dropped out; while in a first- (second-) price sealed-bid auction the high bidder wins the item and pays the highest (second-highest) bid.

An experimental session typically consists of 20–40 auction periods under a given auction institution. Subjects' valuations are determined randomly prior to each auction period (by the experimenter) and are private information. Valuations are typically independent and identical draws (i.i.d) from a *uniform* distribution. In each period the high bidder earns a profit equal to his value less the auction price; other bidders earn zero profit. Bids are commonly restricted to be non-negative and rounded to the nearest penny. Theory does not specify what information feedback bidders ought to get after each auction. Although such information is unimportant in a one-shot auction, it may be important, even critical, to learning given that experimental sessions typically consist of a number of auction periods. Information feedback usually differs between different experimenters, with almost all experimenters reporting back the auction price to all bidders and own earnings to the winning bidder.

Strategic equivalence usually fails between the relevant auction formats: Coppinger et al. (1980) and Cox et al. (1982) found higher prices in first-price than in Dutch auctions (about five per cent higher) with these differences holding across auctions with different numbers of bidders. Further, bidding was significantly above the risk-neutral Nash equilibrium (RNNE) in the first-price auctions for all numbers of bidders $n > 3$, which is consistent with risk- averse bidders.

Kagel et al. (1987) reported failures of strategic equivalence in second-price and English clock auctions, with winning bids in the second-price auctions averaging 11% *above* the predicted equilibrium price. In contrast, market prices converge rapidly to the predicted equilibrium in the clock auctions. Bidding above value in second-price auctions is widespread, with 62% of all bids above values, 30% of all bids essentially equal to value (within five cents of it), and 8% of all bids below it (Kagel and Levin 1993). (In clock auctions price rises by fixed increments with bidders counted as active until they drop out – and are not permitted to re-enter the auction. This format insures clear information flows as a consequence of announcing irrevocable drop-out prices.)

Bidding above value in second-price auctions is attributable to a number of factors: (*a*) it is sustainable since average profits are positive, (*b*) figuring out the dominant strategy is not that obvious, and (*c*) the feedback from losses that would promote the dominant bidding strategy is weak (Kagel et al. 1987). Subsequent research generalizes the superiority of the (dynamic) clock auction format compared to the (static) sealed-bid format to Vickrey-style auctions in which bidders demand multiple units. The closer conformity to equilibrium outcomes in the clock auctions results from the clock format in conjunction with bidders knowing that the auction ends when the next-to-last bidder drops out. This induces bidders to remain active as long as the clock price is less than their value (as they have nothing to lose by remaining active and might win the item) and to

drop out once the price is greater than their value (as they will lose money for sure should they win the item) (Kagel and Levin 2006).

Efficiency in private value auctions can be measured by the percentage of auctions won by the high-value holder. In Cox et al. (1982) 88% of the first-price auctions were Pareto efficient compared with 80% of the Dutch auctions. In contrast, efficiency in first- and second-price auctions may be quite comparable; for example, 82% of the first-price auctions and 79% of the second-price auctions reported in Kagel and Levin (1993) were Pareto efficient. More work needs to be devoted to comparing efficiency across auction institutions.

A number of papers have explored bidding above the RNNE in first-price sealed-bid auctions, questioning the risk-aversion interpretation. This has generated some heated debate (see the December 1992 issue of the *American Economic Review).* Isaac and James (2000) compare estimates of risk preferences from first- price auctions with estimates using the Becker–DeGroot–Marshak (BDM) procedure for comparably risky choices. The Spearman rank–correlation coefficient between individual subject risk parameters is significantly *negatively* correlated under the two procedures. Subjects whose bids in the first-price auction are relatively risk neutral remain risk neutral under BDM, but those who are relatively risk averse in the first- price auction become relatively risk loving under BDM. The net result is that *aggregate* measures of risk preferences show that bidders are risk averse in the first- price auction but risk neutral, or moderately risk loving, under the BDM procedure. Although it is well known from the psychology literature that different elicitation procedures will yield somewhat different quantitative predictions, a negative correlation between measures seems rather astonishing. (See Dorsey and Razzolini 2003, for a similar investigation.)

Neugebauer and Selten (2006) compare treatments with different information feedback: (i) a bidder only learns if s/he won the auction or not, (ii) the winning bid (market price) is revealed to bidders whether they win or not; and (iii) the winning bid is revealed to bidders and the winner learns the second highest bid as well. They find that average bids are highest under treatment (ii) and exceed the RNNE for every given market size. In contrast, bidding above the RNNE does not occur consistently, or is not as strong, in the other two treatments. They use 'learning direction theory' to argue that the information feedback in (ii) promotes bidding above the RNNE. However, the result for treatment (iii) contrasts with results from Kagel et al. (1987) and Dyer et al. (1989a), who find consistent bidding above the RNNE when providing bidders with all bids and valuations following each auction. Perhaps the best conclusion at this point is that subjects typically act 'as if' they are risk averse in first-price auctions, while the underlying basis of their behaviour remains open to interpretation.

In spite of the reported deviations from equilibrium outcomes reported above, the comparative static implications of the IPV model tend to hold (albeit with varying levels of noise). Bidding in first-price auctions increases regularly in response to increased numbers of bidders. For example, in a series of first-price sealed-bid auctions, 86% of subjects increased their bids when the number of bidders increased from five to ten, with the majority of these increases (60%) being statistically significant, with no subjects decreasing their bids by a statistically significant amount (Battalio et al. 1990. More aggressive bidding in response to increased numbers of rivals would seem to be a natural reaction, and can be rationalized by plausible ad hoc rules of thumb.

Kagel and Levin (1993) provide a more stringent test of the comparative static implications of the IPV model using a third-price auction in which the high bidder wins the item and pays the third-highest bid. In this case the model predicts that bids will be above values and will be *reduced* in response to increases in n. They find that 85–90% of all bids are above value compared with 58–67% in second-price auctions and less than 0.5% in first-price auctions. Further, comparing auctions with $n = 5$ and $n = 10$ (i) in first-price auctions *all* bidders increased their bids on average (average increase of $0.65 per auction; $p < .01$), (ii) in second-price auctions the majority of bidders did not change their bids on average

(average decrease of \$0.04; $p > .10$), and (iii) in third-price auctions 46% of all subjects *decreased* their bids on average (average decrease of \$0.40 per auction; $p < .05$). Even stronger qualitative support for the theory is reported when the calculations are restricted to valuations lying in the top half of the domain of valuations (where bidders have a realistic chance of winning and might be expected to take bidding more seriously). Thus, although a number of bidders in third-price auctions clearly err in response to increased numbers of rivals by increasing, or not changing, their bids, the change in pricing rules has relatively large and statistically significant effects on bidders' responses in the *direction* that Nash equilibrium bidding theory predicts. This experiment also illustrates one of the great strengths of the experimental method as there are no third-price auctions outside the lab, where it was developed for the explicit purpose of providing unusual, counter-intuitive predictions to use in testing the theory. The results are increased confidence in the fundamental 'gravitational' forces underlying the theory, in spite of violations of its point predictions. The latter could be the result of some uncontrolled factor impacting on behaviour and/or simple miscalibration on subjects' part.

Single-Unit Common Value Auctions

In common value auctions (CVA) the value of the item is the same to all bidders. What makes common value auctions interesting is that bidders receive signals (estimates) that are correlated (affiliated) with the value of the item but they do not know its true value. Mineral rights auctions (for example, outer continental shelf – OCS – oil lease auctions) are usually modelled as a common value auction. There is a common value element to most auctions. Bidders for a painting may purchase it for their own pleasure, a private value element, but also for investment and eventual resale, the common value element.

Experimental research on CVAs has focused on the 'winner's curse'. Although all bidders obtain unbiased estimates of the item's value, they typically win in cases where they have (one of) the highest signal value. Unless this adverse selection problem is accounted for, it will result in winning bids that are systematically too high, earning below normal or negative profits – a disequilibrium phenomenon. Oil companies claim they fell prey to the winner's curse in early OCS lease sales, with similar claims made in a variety of other settings (for example, free agency markets for professional athletes and corporate takeovers). Economists are naturally sceptical of such claims as they involve out-of-equilibrium play. Experiments clearly show the presence of a winner's curse for inexperienced bidders under a variety of circumstances and with different experimental subjects: average undergraduate or MBA students (Bazeramn and Samuelson 1983; Kagel and Levin 1986), extremely bright (Cal Tech) undergraduates (Lind and Plott 1991), experienced professionals in a laboratory setting (Dyer et al. 1989b), and auctions in which it is common knowledge that one bidder knows, with certainty, the value of the item (Kagel and Levin 1999). Further, these deviations from equilibrium predictions cannot be explained by simple miscalibration on bidders' part as the theory's comparative static implications are systematically violated when bidders suffer from a winner's curse; for example, bidder responses to additional information or increased numbers of rivals.

Kagel et al. (1989) find that inexperienced bidders suffer a pervasive winner's curse in first-price, sealed-bid auctions. For the first nine auctions, profits averaged minus \$2.57 compared with the RNNE prediction of \$1.90, with only 17% of all auctions having positive profits. This is not a simple matter of bad luck as 59% of all bids, and 82% of the high bids, were above the expected value of the item conditional on winning the auction. Although public information in first-price auctions is predicted to raise sellers' revenue, it reduces it for inexperienced bidders as subjects use the public information to help overcome the winner's curse (Kagel and Levin 1986). Similarly, 'public information' reduces revenue in English clock auctions when bidders suffer from a winner's curse (Levin et al. 1996). Further, experienced bidders appear to adjust to the winner's curse through a 'hot stove' learning process: with

the losses, bids are lowered and losses are mitigated, or eliminated, but there is no real understanding of the adverse selection problem. For example, an increase in n generates higher individual bids, although theory predicts a slight reduction (Kagel and Levin 1986). Efforts to explain the winner's curse in terms of limited liability for losses and/or the 'joy of winning' fail as well (Kagel and Levin 1991; Holt and Sherman 1994). In short, inexperienced subjects do not perform well in pure common value auctions.

Experienced subjects learn to overcome the worst effects of the winner's curse, earning positive average profits. But these rarely exceed 65% of the RNNE profit, and virtually all subjects are *not* best responding to their rivals' overly aggressive bids (Kagel and Richard 2001). However, once bidders overcome the worst effects of the winner's curse, public information raises sellers' revenue, English auctions raise more revenue than sealed-bid auctions, and a number of other comparative static implications of the theory are satisfied as well (Kagel and Levin 2002). Experienced bidders learn to overcome the winner's curse through a combination of individual learning and market selection process whereby bankrupt bidders self-select out of further experimental sessions. Ability as measured by composite SAT/ACT scores (standardized college entrance exam scores) matters in terms of avoiding the winner's curse, with the biggest and most consistent impact resulting from those with *below median* scores being more susceptible to the winner's curse. Economics and business majors consistently bid more aggressively than others (thus, lose more), and women, at least initially, are much more susceptible to a winner's curse than men. However, there is still a winner's curse even for the best-calibrated demographic and ability groups (Casari et al. 2007).

Experiments Combining Common-Value and Private-Value Elements

Goeree and Offerman (2002) provide the only experimental study to date in which the object's expected value depends on both private and common value elements. (The difficulty here is in combining private and common value information into a single statistic that maps into a bid.) Actual bids lie in between the RNNE benchmark of fully rational bidding and the naive benchmark in which subjects completely fail to account for the winner's curse. The winner's curse effect is more pronounced the less important a bidder's private value is relative to the common value. Realized efficiency is roughly at the level predicted under the RNNE, with the winner's curse only raising seller revenue and cutting into bidder profits. This occurs because (*a*) almost all bidders suffer from a winner's curse and (*b*) the degree of suffering is roughly the same across bidders, so that the size of the private value element serves to dictate who wins the item.

In an almost common value auction one bidder, the advantaged bidder, has an added private value for the item, unlike all the other (regular) bidders who care only about the common value. With only two bidders, even a tiny private value advantage is predicted to have an explosive effect in second-price sealed-bid auctions: the advantaged bidder always wins and revenue decreases dramatically as the regular bidder lowers her bid to protect against a winner's curse. This effect extends to a variety of English auctions that start with more than two bidders, raising serious concerns about the English auction format (Klemperer 1998). Three experiments have looked at almost common value auctions using both second-price sealed-bid and clock auctions (Avery and Kagel 1997; Rose and Levin 2005; and Rose and Kagel 2005). In all cases the response to the private value advantage has been proportional rather than explosive. This is true even with experienced bidders who earn a respectable share of RNNE profits in pure common value first-price and clock auctions (Rose and Kagel 2005). The apparent reason for these failures is that bidders do not fully appreciate the adverse selection effect conditional on winning, which is exacerbated for regular bidders with an advantaged rival. As such, the behavioural mechanism underlying the explosive effect is not present, and there are no forces at work to replace it.

Internet Auctions

Internet auctions provide new opportunities to conduct experiments to study old and new puzzles. Lucking-Reiley (1999) has used the Internet to sell collectable trading cards under the four standard auction formats, testing the revenue equivalence theorem. He finds that Dutch auctions produce 30% higher revenue than first- price auctions, a reversal of previous laboratory results, and that English and second- price auctions produce roughly equivalent revenue. These results are interesting but lack the controls present in more standard laboratory experiments; that is, there may well be a common value element to the trading cards, and Dutch auctions provide an opportunity to use the game cards immediately, which cannot be done until the fixed closing date in the first-price auctions. Garratt et al. (2004) conduct a second-price auction, recruiting subjects with substantial experience bidding on eBay. Using induced valuations, they find that average bids are close to valuations, but those with prior experience as *sellers* tend to underbid and those with prior experience as *buyers* tend to overbid.

In eBay auctions which have a fixed closing time many bidders *snipe* (submit bids seconds before the closing time), while other bidders increase their bids over time in response to higher bids. This seems puzzling since eBay has a number of characteristics similar to a second-price auction. In addition, there is substantially more last-minute bidding for comparable (private-value) items in eBay than in Amazon auctions, which automatically extend the deadline in response to last-minute bids. Roth and Ockenfels (2002) argue that sniping results from the fixed deadline in eBay, suggesting at least two rational reasons for sniping. Because there are differences between eBay and Amazon other than their ending rules, they conduct a laboratory experiment in which the only difference between auction institutions is the ending rule – a dynamic eBay auction with a .8 (1.0) probability that a late bid will be accepted (eBay.8 and eBay1, respectively) and an Amazon-style auction with a .8 probability that a late bid will be accepted, in which case the auction is automatically extended (Ariely et al. 2005). The results show quite clearly that there is more late bidding in both eBay auctions than in the Amazon auction. Further, there is significantly more late bidding in eBay1 than in eBay.8, which at least rules out one possible rational explanation for sniping – implicit collusion on the part of snipers in an effort to get the item at rock-bottom prices since not all last- minute bids will be recorded (due to congestion) at the website.

Salmon and Wilson (2008) investigate the Internet practice of second-chance offers to non-winning bidders when selling multiple (identical) items. They compare a two-stage game with a second-price auction followed by an ultimatum game between the seller and the second-highest bidder with a sequential English auction. As predicted, the auction-ultimatum game mechanism generates more revenue than the sequential English auction.

Multi-Unit Demand Auctions

Most of the work on multi-unit demand auctions has been devoted to mechanism design issues, in particular dealing with problems created by complementarities, or synergies, between items. Absent package bidding, the latter can create an 'exposure' problem whereby efficient outcomes require submitting bids above the stand-alone values for individual units since the value of the package is more than the sum of the individual values. Correcting for this problem by permitting package bids increases the complexity of the auction significantly, and creates a 'threshold' problem whereby 'small' bidders (for example, those with only local markets) could, in combination, potentially outbid a large competitor who can internalize the complementarities. But the small bidders have no means to coordinate their bids. Leading examples of this line of research are Porter et al. (2003), Kwasnica et al. (2005), and Goeree et al. (2006). Much more work remains to be done in this area.

See Also

- ▶ Auctions (Applications)
- ▶ Auctions (Empirics)
- ▶ Auctions (Theory)

Bibliography

Ariely, D., A. Ockenfels, and A.E. Roth. 2005. An experimental analysis of ending rules in internet auctions. *RAND Journal of Economics* 36: 890–907.

Avery, C., and J.H. Kagel. 1997. Second-price auctions with asymmetric payoffs: An experimental investigation. *Journal of Economics and Management Strategy* 6: 573–604.

Battalio, R.C., C.A. Kogut, and D.J. Meyer. 1990. Individual and market bidding in a Vickrey first-price auction: Varying market size and information. In *Advances in behavioral economics*, vol. 2, ed. L. Green and J.H. Kagel. Norwood: Alex.

Bazeramn, M.H., and W.F. Samuelson. 1983. I won the auction but don't want the prize. *Journal of Conflict Resolution* 27: 618–634.

Casari, M., J.C. Ham, and J.H. Kagel. 2007. Selection bias, demographic effects and ability effects in common value auction experiments. *American Economic Review* 97: 1278–1304.

Coppinger, V.M., V.L. Smith, and J.A. Titus. 1980. Incentives and behavior in English, Dutch and sealed-bid auctions. *Economic Inquiry* 43: 1–22.

Cox, J., B. Roberson, and V.L. Smith. 1982. Theory and behavior of single object auctions. In *Research in experimental economics*, ed. V.L. Smith. Greenwich: JAI Press.

Dorsey, R., and L. Razzolini. 2003. Explaining overbidding in first price auctions using controlled lotteries. *Experimental Economics* 6: 123–140.

Dyer, D., J.H. Kagel, and D. Levin. 1989a. Resolving uncertainty about the number of bidders in independent private-value auctions: An experimental analysis. *RAND Journal of Economics* 20: 268–279.

Dyer, D., J.H. Kagel, and D. Levin. 1989b. A comparison of naive and experienced bidders in common value offer auctions: A laboratory analysis. *Economic Journal* 99: 108–115.

Garratt, R., M. Walker, and J. Wooders. 2004. *Behavior in second-price auctions by highly experienced eBay buyers and sellers*, Working paper, vol. 1181. Santa Barbara: Department of Economics, UC.

Goeree, J.K., and T. Offerman. 2002. Efficiency in auctions with private and common values: An experimental study. *American Economic Review* 92: 625–643.

Goeree, J.K., C.A. Holt, and J. O. Ledyard. 2006. *An experimental comparison of the FCC's combinatorial and non-combinatorial simultaneous multiple round auctions*. Prepared for the wireless telecommunications bureau of the federal communications commission. Online. Available at http://wireless.fcc.gov/auctions/data/papersAndStudies/fcc_final_report_071206.pdf. Accessed 1 Feb 2007.

Holt Jr., C.A., and R. Sherman. 1994. The loser's curse and bidder's bias. *American Economic Review* 84: 642–652.

Isaac, M., and D. James. 2000. Just who are you calling risk averse? *Journal of Risk and Uncertainty* 20: 177–187.

Kagel, J.H., and D. Levin. 1986. The winner's curse and public information in common value auctions. *American Economic Review* 76: 894–920.

Kagel, J.H., and D. Levin. 1991. The winner's curse and public information in common value auctions: Reply. *American Economic Review* 81: 362–369.

Kagel, J.H., and D. Levin. 1993. Independent private value auctions: Bidder behavior in first-, second- and third-price auctions with varying numbers of bidders. *Economic Journal* 103: 868–879.

Kagel, J.H., and D. Levin. 1999. Common value auctions with insider information. *Econometrica* 67: 1219–1238.

Kagel, J.H., and D. Levin. 2001. Behavior in multi-unit demand auctions: Experiments with uniform price and dynamic Vickrey auctions. *Econometrica* 69: 413–454.

Kagel, J.H., and D. Levin. 2002. Bidding in common value auctions: A survey of experimental research. In *Common value auctions and the Winner's curse*. Princeton: Princeton University Press.

Kagel, J.H. and D. Levin. 2006. Implementing efficient multi-object auction institutions: An experimental study of the performance of boundedly rational agents. Mimeo, Ohio State University.

Kagel, J.H., and J.F. Richard. 2001. Super-experienced bidders in first-price common value auctions: rules of thumb, Nash equilibrium bidding and the winner's curse. *Review of Economics and Statistics* 83: 408–419.

Kagel, J.H., R.M. Harstad, and D. Levin. 1987. Information impact and allocation rules in auctions with affiliated private values: A laboratory study. *Econometrica* 55: 1275–1304.

Kagel, J.H., D. Levin, R. Battalio, and D.J. Meyer. 1989. First-price common value auctions: Bidder behavior and the winner's curse. *Economic Inquiry* 27: 241–258.

Klemperer, P. 1998. Auctions with almost common values: The 'wallet game' and its applications. *European Economic Review* 42: 757–769.

Kwasnica, A.M., J.O. Ledyard, D. Porter, and C. DeMartini. 2005. A new and improved design for multiobject iterative auctions. *Management Science* 51: 419–434.

Levin, D., J.H. Kagel, and J.F. Richard. 1996. Revenue effects and information processing in English common value auctions. *American Economic Review* 86: 442–460.

Lind, B., and C.R. Plott. 1991. The winner's curse: Experiments with buyers and with sellers. *American Economic Review* 81: 335–346.

Lucking-Reiley, D. 1999. Using field experiments to test equivalence between auction formats: Magic on the internet. *American Economic Review* 89: 1062–1080.

Neugebauer, T., and R. Selten. 2006. Individual behavior of first-price auctions: the importance of information feedback in computerized experimental markets. *Games and Economic Behavior* 54: 183–204.

Porter, D., S. Rassenti, A. Roopnarine, and V. Smith. 2003. Combinatorial auction design. *Proceedings of the National Academy of Sciences* 100: 11153–11157.

Rose, S.L. and J.K Kagel. 2005. Bidding in almost common value auctions: An experiment, mimeographed. Mimeo, Ohio State University.
Rose, S.L. and D. Levin. 2005. An experimental investigation of the explosive effect in common value auctions. Mimeo, Ohio State University.
Roth, A.E., and A. Ockenfels. 2002. Last-minute bidding and the rules for ending second-price auctions: evidence from eBay and Amazon auctions on the internet. *American Economic Review* 92: 1093–1103.
Salmon, T.C., and B.J. Wilson. 2008. Second chance offers versus sequential auctions: Theory and behavior. *Economic Theory* 34: 47–67.

compatibility; Linkage principle; Optimal auctions; Pay-as-bid auction; Private information; Revenue equivalence theorem; Sealed-bid auctions; Second-price auction; Simultaneous ascending auctions; Sincere bidding; Spectrum auctions; Subgame perfection; Substitutes; Tâtonnement; Treasury auctions; Uniform-price auction; Vickrey auction; Vickrey, W. S.; Walrasian auctioneer; Winner's curse

JEL Classifications
D44

Auctions (Theory)

Lawrence M. Ausubel

Abstract

Auction theory has undergone two waves of innovation. The first, which originated with Vickrey (1961) and was completed in the early 1980s, focused on single-item auctions. Results included: guiding principles such as revenue equivalence; the derivation of the optimal auction; and comparisons of first-price, second-price and English auctions. The second, influenced by Treasury and spectrum auctions, emerged in the 1990s and dealt particularly with multi-item auctions. Research has studied: static auctions, including pay-as-bid and uniform-price auctions; dynamic auctions such as simultaneous ascending and clock auctions; combinatorial auctions; and efficient auction design. Much progress has been made, but outstanding problems remain.

Keywords

Auctions; Auctions (theory); Bidding; Clock auction; Coalitional form game; Combinatorial auctions; Complements; Core; Demand reduction; Dutch auction; Efficient auctions; English auction; Envelope theorem; Equilibrium; Experimental economics; First-price auction; Game theory; Games of complete information; Games of incomplete information; Incentive

Auctions occupy a deservedly prominent place within microeconomics and game theory, for at least three reasons.

First, the auction is, in its own right, an important device for trade. Auctions have long been a common way of selling diverse items such as works of art and government securities. In recent years, their importance in consumer markets has increased through the ascendancy of eBay and other Internet auctions. At the same time, the use of auctions for transactions between businesses has expanded greatly, most notably in the telecommunications, energy and environmental sectors, and for procurement purposes generally.

Second, auctions have become the clearest success story in the application of game theory to economics. In most applications of game theory, the modeller has considerable (perhaps excessive) freedom to formulate the rules of the game, and the results obtained will often be highly sensitive to the chosen formulation. By way of contrast, an auction will typically have a well-defined set of rules, yielding clearer theoretical predictions.

Third, there has been an increasing wealth of auction data available for empirical analysis in recent years. In conjunction with the available theory, this has led to a growing body of empirical work on auctions. Moreover, auctions are very well suited for laboratory experiments and they have been a very fruitful area for experimental economics.

This article is limited in its scope to *auction theory.* Other related articles, reviewing empirical and experimental work on auctions and the

theoretical analysis of mechanism design, are cross-referenced at the end.

Introduction

Auction theory is often said to have originated in the seminal 1961 article by William Vickrey. While Vickrey's insights were initially unrecognized and it would be many years before his work was followed up by other researchers, it eventually led to a formidable body of research by pioneers including Wilson, Clarke, Groves, Milgrom, Weber, Myerson, Maskin and Riley. The first wave of theoretical research into auctions was concluded in the mid-1980s, by which time there was a widespread sense that it had become a relatively complete body of work with very little remaining to be discovered. See McAfee and McMillan (1987) for an excellent review of the first wave of auction theory.

However, the perception that auction theory was complete began to change following two pivotal events in the 1990s: the Salomon Brothers scandal in the US government securities market in 1991, and the advent of the Federal Communications Commission (FCC) spectrum auctions in 1994. In the aftermath of the former, the Department of the Treasury sought input from academia concerning the US Treasury auctions. In the preparation for the latter, the FCC encouraged the active involvement of auction theorists in the design of the new auctions.

Each of these two episodes undoubtedly benefitted from the participation of academics. In particular, the FCC introduced an innovative dynamic auction format – the simultaneous ascending auction – whose empirical performance appears far superior to previous static sealed-bid auctions. The Treasury's experimentation with, and eventual adoption of, uniform-price auctions in place of pay-as-bid auctions also appears to have resulted from economists' input.

At the same time, these two pivotal events underscored some extremely serious limitations in auction theory as it existed in the early to mid-1990s. It became apparent then that the theory that had been developed was almost exclusively one of single-item auctions, and that relatively little was established concerning multi-item auctions. As the flip side of the same coin, these episodes made it obvious that many of the empirically important examples of auctions involve a multiplicity of items. As a result, a second wave of theoretical research into auctions, focusing especially on multi-item auctions, emerged in the middle of the 1990s and continued into the 21st century.

This article begins by reviewing the theory of single-item auctions, largely completed during the first period of research. It continues by reviewing the theory of multi-unit auctions, still a work in progress as of 2007.

The scope and detail of the present article is necessarily quite limited. For deeper and more comprehensive treatments of auctions, three notable books, by Krishna (2002), Milgrom (2004) and Cramton et al. (2006), are especially recommended to readers. Earlier survey articles by McAfee and McMillan (1987) and Wilson (1992) also provide excellent treatments of the literature on single-item auctions. A compendium by Klemperer (2000) brings together many of the best articles in auction theory.

Sealed-Bid Auctions for Single Items

Much of the analysis within traditional auction theory has concerned sealed-bid auctions (that is, static games) for single items. Bidders submit their sealed bids in advance of a deadline, without knowledge of any of their opponents' bids. After the deadline, the auctioneer unseals the bids and determines a winner. The following are the two most commonly studied sealed-bid formats:

- *First-price auction*: the highest bidder wins the item, and pays the amount of his bid.
- *Second-price auction*: the highest bidder wins the item, and pays the amount bid by the second-highest bidder.

Note that the above auction formats (and, indeed, all of the auctions described in this article) have been described for a regular auction in which the auctioneer offers items for sale and the bidders

are buyers. Each can easily be restated for a 'reverse auction' (that is, procurement auction) in which the auctioneer solicits the purchase of items and the bidders are sellers. For example, in a second-price reverse auction, the lowest bidder is chosen to provide the item and is paid the amount bid by the second-lowest bidder.

The Private Values Model

A seller wishes to allocate a single unit of a good or service among n bidders ($i = 1, \ldots, n$). The bidders bid simultaneously and independently as in a non-cooperative static game. Bidder i's payoff from receiving the item in return for the payment y is given by $v_i - y$ (whereas bidder i's payoff from not winning the item is normalized to zero). Each bidder i's valuation, v_i, for the item is private information. Bidder i knows v_i at the time he submits his bid. Meanwhile, the opposing bidders $j \neq i$ view v_i as a random variable whose realization is unknown, but which is drawn according to the known joint distribution function $\hat{F}(v_1, \ldots, v_i, \ldots, v_n)$.

This model is referred to as the *private values* model, on account that each bidder's valuation depends only on his own – and not the other bidders' – information. (By contrast, in a *pure common values* model, $v_i = v_j$, for all $i, j = 1, \ldots, n$; and in an *interdependent values* model, bidder i's valuation is allowed to be a function of $v_{-i} = \{v_j\}_{j \neq i}$, as well as of v_i.) With private values, some especially simple and elegant results hold, particularly for the second-price auction.

Two additional assumptions are frequently made. First, we generally assume that bidders are *risk neutral* in evaluating their payoffs under uncertainty. That is, each bidder seeks merely to maximize the mathematical expectation of his payoff. Second, we often assume *independence* of the private information. That is, the joint distribution function, $\hat{F}(v_1, \ldots v_n)$, is given by the product of separate distribution functions, $F_i(\cdot)$, for each of the v_i. However, both the risk neutrality and independence assumptions are unnecessary for solving the second-price auction, which we analyse first.

Solution of the Second-Price Auction

Sincere bidding (that is, the truthful bidding of one's own valuation) is a Nash equilibrium of the sealed-bid second-price auction, under private values. That is, if each bidder i submits the bid $b_i = v_i$, then there is no incentive for any bidder to unilaterally deviate. Moreover, sincere bidding is a weakly dominant strategy for each bidder; and sincere bidding by all bidders is the unique outcome of elimination of weakly dominated strategies. These facts make the sincere bidding equilibrium an especially compelling outcome of the second-price auction.

Let $\hat{b}_{-i} = \max_{j \neq i}\{b_j\}$, the highest among the opponents' bids. The dominant strategy property is easily established by comparing bidder i's payoff from the sincere bid of $b_i = v_i$ with his payoff from instead bidding $b_i' < v_i$ ('shading' his bid). If $\hat{b}_{-i}$ is less than b_i' or greater than v_i then bid-shading has no effect on bidder i's payoff; in the former case, bidder i wins either way, and in the latter case, bidder i loses either way. However, in the event that $\hat{b}_{-i}$ is between b_i' and v_i, the bid-shading makes a difference: if bidder i bids v_i, he wins the auction and thereby achieves a positive payoff of $v_i - \hat{b}_{-i} > 0$; whereas, if bidder i bids b_i', he loses the auction and receives zero payoff. Thus, $b_i = v_i$ weakly dominates any bid $b_i' > v_i$. A similar comparison finds that $b_i = v_i$ weakly dominates any bid $b_i' > v_i$. Sincere bidding is optimal, regardless of the bidding strategies of opposing bidders.

Note that the above argument in no way uses the risk neutrality or independence assumptions, nor does it require any form of symmetry. Sincere bidding may also be viewed as an *ex post equilibrium* of the second-price auction, in the sense that the strategy would remain optimal even if the bidder were to learn his opponents' bids before he was required to submit his own bid. Indeed, one of the strengths of the result that sincere bidding is a Nash equilibrium in weakly dominant strategies is that it basically relies only upon the private values assumption, and is otherwise extremely robust to the specification of the model.

Incentive Compatibility in Any Sealed-Bid Auction Format

Consider any equilibrium of *any* sealed-bid auction format, in the private values model. Given that bidder i's valuation is private information, observe

that there is nothing to force bidder i to bid according to his true valuation v_i instead of some other valuation w_i. As a result, the equilibrium must have a structure that gives bidder i the incentive to bid according to his true valuation. This requirement is known as *incentive compatibility.*

In the following derivation, we assume that the support of each bidder i's valuation is the interval $[\underline{v}_i, \overline{v}_i]$. We will make both the risk neutrality and independence assumptions. Let $\Pi_i(v_i)$ denote bidder i's expected payoff, let $P_i(v_i)$ denote bidder i's probability of winning the item, and let $Q_i(v_i)$ denote bidder i's expected payment in this equilibrium, when his valuation is v_i. The reader should note that $Q_i(v_i)$ refers here to bidder i's unconditional expected payment, *not* to his expected payment conditional on winning. Given the risk-neutrality assumption, $\Pi_i(v_i)$ is given by:

$$\Pi_i(v_i) = P_i(v_i)v_i - Q_i(v_i). \tag{1}$$

Next, we pursue the observation that there is nothing forcing bidder i to bid according to his true valuation v_i rather than according to another valuation w_i. Define $\pi_i(w_i, v_i)$ to be bidder i's expected payoff from employing the bidding strategy of a bidder with valuation w_i when his true valuation is v_i. Observe that:

$$\pi_i(w_i, v_i) = P_i(w_i)v_i - Q_i(w_i), \tag{2}$$

since bidder i's probability of winning and expected payment depend exclusively on his bid, not on his true valuation. Bidder i will voluntarily choose to bid according to his true valuation only if his expected payoff is greater than from bidding according to another valuation w_i, that is, if:

$$\Pi_i(v_i) \geq \pi_i(w_i, v_i), \quad \text{for all } v_i,\ w_i \in [\underline{v}_i, \overline{v}_i] \text{ and all } i = 1, \ldots, n. \tag{3}$$

Inequality (3), referred to as the *incentive-compatibility constraint,* has very strong implications.

Next, note that $\Pi_i(v_i) = \pi_i(v_i, v_i) = \max_{w_i \in [\underline{v}_i, \overline{v}_i]} \pi_i(w_i, v_i)$. It is straightforward to see that $\Pi_i(\cdot)$ is monotonically non-decreasing and continuous. Consequently, it is differentiable almost everywhere and equals the integral of its derivative. Applying the envelope theorem at any v_t where $\Pi_i(\cdot)$ is differentiable yields:

$$\frac{d\Pi_i(v_i)}{dv_i} = \frac{\partial \pi_i(w_i,\ v_i)}{\partial v_i}\Big|_{w_i = v_i} = P_i(w_i)\big|_{w_i = v_i} = P_i(v_i). \tag{4}$$

Integrating Eq. (4), we have:

$$\Pi_i(v_i) = \Pi_i(\underline{v}_i) + \int_{\underline{v}_i}^{v_i} P_i(x)dx, \quad \text{for all } v_i \in [\underline{v}_i, \overline{v}_i] \text{ and all } i = 1, \ldots, n. \tag{5}$$

Solution of the First-Price Auction

The sealed-bid first-price auction requires two symmetry assumptions in order to yield a fairly simple solution. First, we assume *symmetric bidders,* in the sense that the joint distribution function $\hat{F}(v_1, \ldots, v_i, \ldots, v_n)$ governing the bidders' valuations is a symmetric function of its arguments. This assumption and the associated notation are simplest to state if independence is assumed. In this case, we write $F_i(\cdot)$ for the distribution function of each v_i; symmetry is the assumption that $F_i = F$, for all $i = 1,\ldots, n$, or, in other words, the assumption that the various v_i are identically distributed, as well as independent, random variables. However, a similar derivation with only slightly more cumbersome notation is possible if the bidders are symmetric but the v_i are affiliated random variables. We write $[\underline{v}, \overline{v}]$ for the support of $F(\cdot)$. In addition, we assume that $F(\cdot)$ is a continuous function, so that there are no mass points in the common probability distribution of the bidders' valuations.

Second, we restrict attention to *symmetric, monotonically increasing equilibria* in pure strategies. The assumed symmetry of bidders opens the possibility for existence of a symmetric equilibrium. (Meanwhile, asymmetric equilibria are also possible in symmetric games, but Maskin and Riley 2003, establish that, under slightly

stronger assumptions, the construction here gives the unique equilibrium of the auction.) Any pure-strategy equilibrium can be characterized by the bid functions $\{B_i(\cdot)\}_{i=1}^n$, which give bidder i's bid $B_i(v_i)$ when his valuation is v_i. Our assumption is that $B_t = B$, for all $i = 1, \ldots, n$, where $B(\cdot)$ is a strictly increasing function.

Observe that, in any symmetric equilibrium, bidder i wins against bidder j if and only if $B(v_j) < B(v_i)$ and, given strict monotonicity, if and only if $v_j < v_i$. (We can ignore the event $v_j = v_i$; this is a zero-probability event, since we have assumed the distribution of valuations has no mass points.) Consequently, bidder i wins the item if and only $v_j < v_i$ for all $j \neq i$. Since the $\{v_j\}_{j \neq i}$ are i.i.d. random variables, bidder i has probability $F(v_i)^{n-1}$ of winning the auction when his valuation is v_i. We write: $P_i(v_i) = F(v_i)^{n-1}$, for all $v_i \in [\underline{v}, \overline{v}]$ and all $i = 1, \ldots, n$.

Moreover, in a first-price auction, the bidder's payoff equals $v_i - B(v_i)$ if he wins the auction and zero if he loses. Consequently his expected payoff equals:

$$\begin{aligned}\Pi_i(v_i) &= P_i(v_i)[v_i - B(v_i)] \\ &= F(v_i)^{n-1}[v_i - B(v_i)]. \end{aligned} \tag{6}$$

Observe from Eq. (6) that, if $v_i = \underline{v}$, bidder i's probability of winning equals zero and, hence, $\Pi_i(\underline{v}) = 0$. Substituting this fact and $P_i(v_i) = F(v_i)^{n-1}$ into Eq. (5) yields:

$$\Pi_i(v_i) = \int_{\underline{v}}^{v_i} F(x)^{n-1} dx, \quad \text{for all } v_i \in [\underline{v}, \overline{v}] \text{ and all } i = 1, \ldots, n. \tag{7}$$

Combining Eq. (6) with Eq. (7), and solving for B(·), yields the equilibrium bid function:

$$B(v_i) = v_i - \frac{\Pi_i(v_i)}{F(v_i)^{n-1}} = v_i - \frac{\int_{\underline{v}}^{v_i} F(x)^{n-1} dx}{F(v_i)^{n-1}}. \tag{8}$$

The posited strict monotonicity is verified by differentiating Eq. (8) with respect to v_u which shows that $B'(v_i) > 0$. Thus, Eq. (8) provides us with the unique symmetric equilibrium in pure strategies of the sealed-bid first-price auction. This result holds for arbitrary continuous distribution functions $F(\cdot)$ with support on an interval $[\underline{v}, \overline{v}]$.

Revenue Equivalence, Efficient Auctions and Optimal Auctions

Standard practice in auction theory is to evaluate auction formats according to either of two criteria: efficiency and revenue optimization. With the quasi-linear utilities generally assumed in auction theory, efficiency means putting the items in the hands of those who value them the most. Revenue maximization means maximizing the seller's expected revenues or, in a procurement auction, minimizing the buyer's expected procurement costs. In auctions of government assets such as spectrum licenses, the explicit objective is often efficiency. In auctions by private parties, the explicit objective is often revenue optimization.

Efficient Auctions

The above solutions to the second-price and first-price auctions both yield full efficiency. In the symmetric increasing equilibrium of the first-price auction, the highest bid corresponds to the highest valuation, and so the item is assigned efficiently for every realization of the random variables. In the dominant strategy equilibrium of the second-price auction, the identical conclusion holds. Thus, in a symmetric private values model, an objective of efficiency looks kindly upon both auction formats – but does not prefer one over the other.

Revenue Equivalence

One of the classic and most far-reaching results in auction theory is revenue equivalence, which provides a set of assumptions under which the sellers' and buyers' expected payoffs are guaranteed to be the same under different auction formats.

Revenue equivalence (Vickrey 1961; Myerson 1981; Riley and Samuelson 1981) may be stated as follows. Assume that the random variables representing the bidders' valuations are independent, and assume that bidders are risk neutral. Consider any two auction formats satisfying

both of the following properties: (*a*) the two auction formats assign the item(s) to the same bidder (s), for every realization of random variables; and (*b*) the two auction formats give the same expected payoff to the lowest valuation type, $\underline{v}_i$, of each bidder *i*. Then each bidder earns the same expected payoff under each of the two auction formats and, consequently, the seller earns the same expected revenues under each of the two auction formats.

For an auction of a single item, the result follows directly from Eq. (5) above. Recall that this equation holds for any equilibrium of *any* sealed-bid auction format. If for every realization of the random variables the two auction formats assign the item to the same bidder, then each bidder's probability, $P_i(\cdot)$, of winning is the same under the two auction formats. If in addition, $\Pi_i(\underline{v}_i)$ is the same under the two auction formats, then Eq. (5) implies that the entire function $\Pi_i(\cdot)$ is the same under the two auction formats. Since this holds for every bidder *i*, and since the expected gains from trade are the same under the two auction formats, it follows from an accounting identity that the seller's expected revenues are also the same under the two auction formats.

One of the most important applications of revenue equivalence is that the above solutions to the second-price and first-price auctions give the seller the same expected revenues (and also give each buyer the same expected payoffs). Revenue equivalence is applicable because, as argued above, the item is assigned efficiently for every realization of the random variables in each of these auction formats. Moreover, when $v_i = \underline{v}$, the expected payoff of bidder *i* equals zero in each of these auction formats. To understand this result, observe that (all other things equal) a bidder in a first-price auction will bid lower than in a second-price auction, since the payment rule is less generous. Expected revenues will be greater in the first-price or the second-price auction depending on whether the highest of a collection of smaller bids or the second-highest of a collection of larger bids is greater in expectation. The revenue equivalence theorem establishes that, in the symmetric private values model, the two effects exactly offset one another.

Optimal Auctions

Another classic result of auction theory is the determination of the auction format that optimizes revenues. This result, known in the literature as the *optimal auction,* is due to Harris and Raviv (1981), Myerson (1981), and Riley and Samuelson (1981). Any possible auction format is considered – the item may be assigned to the bidder who submitted the highest bid (as in the second-price or first-price auction), but it may alternatively be allocated to another bidder, randomized in its allocation, or withheld from sale entirely, depending on the collection of bids submitted. At the outset, this might be viewed as a very complicated problem, since it requires selecting simultaneously the probability of winning and a payment that optimizes revenues. However, by using analysis similar to the treatment of incentive compatibility, above, it can be shown that the expected payment is determined up to a constant by the probability of winning. Consequently, the problem simplifies to determining the probability of each bidder winning (for every realization of the random variables) that optimizes revenues.

For symmetric bidders, each of whose distributions satisfies a regularity condition, a particularly simple characterization of the optimal auction can be obtained. Let $F(\cdot)$ be the distribution function of the valuation v_i of each bidder *i*, let $f(\cdot)$ be the associated density function and suppose that $v_i - \frac{1-F(v_i)}{f(v_i)}$ is strictly increasing in v_i for all $v_i \in [\underline{v},\overline{v}]$. Then the optimal auction assigns the item to the bidder *i* with the highest v_i, if and only if the highest v_i exceeds the reserve valuation *r*, where *r* is defined by $r - \frac{1-F(r)}{f(r)} = v_0$ and where v_0 is the seller's valuation for the item.

In other words, with symmetric bidders, both the second-price and the first-price auctions become optimal auctions, once a reserve price of *r* is inserted.

Full Rent Extraction

The optimal auctions problem can be reconsidered without the independence assumption. However, Crémer and McLean (1985) demonstrate that, if the bidders' private information is correlated, then there exists a mechanism that

enables the seller to extract *all* of the gains from trade. The mechanism includes a procedure for allocating the item efficiently. Superimposed on this, the mechanism provides rewards to bidders if their reports of private information 'agree' with each other, and penalties to bidders if their reports 'disagree' with each other. The amounts of the rewards and penalties – both potentially quite large – are set so as to make the bidders indifferent between participating and not participating in the mechanism. As such, the mechanism enables the seller to extract the entire surplus, including the informational rents that the bidders are able to obtain under the independence assumption. This is referred to as *full rent extraction.*

Crémer and McLean's result may be viewed as fundamentally negative, in that it suggests that the optimal auctions analysis may be of limited relevance. Real-world auction mechanisms appear to be broadly consistent with the predictions of the optimal auctions theory under the independence assumption, but they look nothing like the full rent-extracting mechanisms possible with correlated private information. Given that there are good reasons to believe that bidders' private signals are correlated with one another, it would appear that the optimal auctions analysis does not provide us with great insight into real-world auctions. Some subsequent research has attempted to weaken the extreme conclusion of full rent extraction by positing that bidders have limited liability or by introducing opportunities for auctioneer collusion or cheating, but in many respects these devices appear to be ineffectual patches for an elegant theory (optimal auctions) that suffers from only limited empirical relevance.

Dynamic Auctions for Single Items

The next two formats considered for auctioning single items are dynamic auctions: participants bid sequentially over time and, potentially, learn something about their opponents' bids during the course of the auction. In the first dynamic auction, the price *ascends*; and in the second dynamic auction, the price *descends:*

- *English auction*: bidders dynamically submit successively higher bids for the item. The final bidder wins the item, and pays the amount of his final bid.
- *Dutch auction*: the auctioneer starts at a high price and announces successively lower prices, until some bidder expresses his willingness to purchase the item by bidding. The first bidder to bid wins the item, and pays the current price at the time he bids.

Note that, as in section "Sealed-Bid Auctions for Single Items", each of these auction formats has been described for a regular auction in which the auctioneer offers items for sale, but can easily be restated for a 'reverse auction'. For example, in an English reverse auction the bids would descend rather than ascend, while in a Dutch reverse auction the auctioneer would offer to buy at successively higher prices.

Solution of the Dutch Auction

An insight due to Vickrey (1961) is that the Dutch auction is strategically equivalent to the sealed-bid first-price auction. To see the equivalence, consider the real meaning of a strategy b_i by bidder i in the Dutch auction: 'If no other bidder bids for the item at any price higher than b_i, then I am willing to step in and purchase it at b_i.' Just as in the sealed-bid first-price auction, the bidder i who selects the highest strategy b_i in the Dutch auction wins the item and pays the amount b_i. Furthermore, although the Dutch auction is explicitly dynamic, there is nothing that can happen that would lead any bidder to want to change his strategy while the auction is still running. If strategy b_i was a best response for bidder i evaluated at the starting price p_0, then b_i remains a best response evaluated at any price $p < p_0$, on the assumption that no other bidder has already bid at a price between p_0 and p. Meanwhile, if another bidder has already bid, then there is nothing that bidder i can do; the Dutch auction is over. Hence, any equilibrium of the sealed-bid first-price auction is also an equilibrium of the Dutch auction, and vice versa.

Solution of the English Auction

By way of contrast, some meaningful learning and/or strategic interaction is possible during an English auction, so the outcome is potentially different from the outcome of the sealed-bid second-price auction.

We model the English auction as a 'clock auction': the auctioneer starts at a low price and announces successively higher prices. At every price, each bidder is asked to indicate his willingness to purchase the item. The price continues to rise so long as two or more bidders indicate interest. The auction concludes at the first price such that fewer than two bidders indicate interest, and the item is awarded at the final price. This clock-auction description is used instead of a game where bidders successively announce higher prices, since it yields simpler arguments and clean results.

With pure private values, the reasonable equilibrium of the English auction corresponds to the dominant-strategy equilibrium of the sealed-bid second-price auction. A bidder's strategy designates the price at which he will drop out of the auction (on the assumption that at least one opponent still remains); in equilibrium, the bidder sets his drop-out price equal to his true valuation. However, matters become more complicated in the case of interdependent valuations, where each bidder's valuation depends not only on his own information, v_i, but also on the opposing bidders' information, v_{-i}. We turn to this case next.

The Winner's Curse and Revenues Under Interdependent Values

One of the most celebrated phenomena in auctions is the 'winner's curse'. Whenever a bidder's valuation depends positively on other bidders' information, winning an item in an auction may confer 'bad news' in the sense that it indicates that other bidders possessed adverse information about the item's value. The potential for falling victim to the winner's curse may induce restrained bidding, curtailing the seller's revenues. In turn, some auction formats may produce higher revenues than others, to the extent that they mitigate the winner's curse and thereby make it safe for bidders to bid more aggressively.

The basic intuition, which is often referred to as the 'linkage principle' and is due to Milgrom and Weber (1982), is that the winner's curse is mitigated to the extent that the winner's payment depends on the opposing bidders' information.

Thus, under appropriate assumptions, the second-price auction will yield higher expected revenues than the first-price auction: the price paid by the winner of a second-price auction depends on the information possessed by the highest losing bidder, while the price paid by the winner of a first-price auction depends exclusively on his own information. Moreover, the English auction will yield higher expected revenues than the second-price auction: the price paid by the winner of an English auction may depend on the information possessed by *all* of the losing bidders (who are observed as they drop out), while the price paid by the winner of a (sealed-bid) second-price auction depends only on the information of the highest losing bidder.

These conclusions require an assumption known as 'affiliation', which intuitively means something very close to 'non-negative correlation'. More precisely, let $v = (v_1, \ldots, v_n)$ and $v' = (v'_1, \ldots, v'_n)$ be possible realizations of the n bidders' random variables, and let $f(\cdot, \ldots, \cdot)$ denote the joint density function. Let $v \vee v'$ denote the component-wise maximum of v and v', and let $v \wedge v'$ denote the component-wise minimum. The random variables v and v' are said to be *affiliated* if:

$$f(v \vee v')f(v \wedge v') \geq f(v)f(v'), \quad \text{for all } v,\ v' \in [\underline{v}, \ldots \overline{v}]^n. \tag{9}$$

Affiliation provides that two high realizations or two low realizations of the random variables are at least as likely as one high and one low realization, and so on, meaning something close to non-negative correlation. Independence is included (as a boundary case) in the definition: for independent random variables, the affiliation inequality (9) is satisfied with equality. To obtain strict revenue rankings, the affiliation inequality must hold strictly.

These conclusions also rely on several symmetry assumptions. Bidders are symmetric, the

equilibria considered are symmetric, and each bidder's valuation depends on all of its opponents' information in a symmetric way. Each bidder's valuation increases (weakly) in its own and its opponents' information, and attention is restricted to equilibria in monotonically increasing strategies. As before, each bidder is risk neutral in evaluating its payoff under uncertainty.

These conclusions also rely on a monotonicity assumption: each bidder's valuation increases (weakly) in its own and in the opposing bidders' information. In addition, as before, each bidder is risk-neutral in evaluating its payoff under uncertainty. Furthermore, the two symmetry assumptions of section "Solution of the First-Price Auction" are made: bidders are symmetric in the sense that the joint distribution governing the bidders' information is a symmetric function of its arguments; and attention is restricted to symmetric, monotonically increasing equilibria in pure strategies.

Under these assumptions, the sealed-bid first-price and second-price auctions and the English auction possess symmetric, monotonic equilibria. However, while these equilibria are all efficient, Milgrom and Weber (1982) establish that they may be ranked by revenues: the English auction yields expected revenues greater than or equal to those of the sealed-bid second-price auction, which in turn yields expected revenues greater than or equal to those of the sealed-bid first-price auction. Their theorem provides one of the most powerful results of auction theory, justifying the conventional wisdom that dynamic auctions yield higher revenues than sealed-bid auctions.

Auctions of Homogeneous Goods

Sealed-Bid, Multi-unit Auction Formats

The defining characteristic of a homogeneous good is that each of the M individual items is identical (or a close substitute), so that bids can be expressed in terms of quantities without indicating the identity of the particular good that is desired. Treating goods as homogeneous has the effect of dramatically simplifying the description of the bids that are submitted and the overall auction procedure. This simplification is especially appropriate in treating subject matter such as financial securities or energy products. Any two $10,000 US government bonds with the same interest rate and the same maturity are identical, just as any two megawatts of electricity provided at the same location on the electrical grid at the same time are identical.

There are three principal sealed-bid, multi-unit auction formats for M homogeneous goods. In each of these, a bid comprises an inverse demand function, that is, a (weakly) decreasing function $p_i(q)$, for $q \in [0, M]$, representing the price offered by bidder i for a first, second, and so on, unit of the good. (Note that this notation may be used to treat situations where the good is perfectly divisible, as well as situations where the good is offered in discrete quantities.) The bidders submit bids; the auctioneer then aggregates the bids and determines a clearing price. Each bidder wins the quantity demanded at the clearing price, but his payment varies according to the particular auction format:

- *Pay-as-bid auction.* Each bidder wins the quantity demanded at the clearing price, and pays the amount that he bid for each unit won.
- *Uniform-price auction.* Each bidder wins the quantity demanded at the clearing price, and pays the clearing price for each unit won.
- *Multi-unit Vickrey auction.* Each bidder wins the quantity demanded at the clearing price, and pays the opportunity cost (relative to the bids submitted) for each unit won.

(Pay-as-bid auctions are also known as 'discriminatory auctions' or 'multiple-price auctions'. Uniform-price auctions are often referred to in the financial press as 'Dutch auctions', generating some confusion with respect to the standard usage of the auction theory literature. They are also known as 'nondiscriminatory auctions', 'competitive auctions' or 'single-price auctions'.)

Sealed-bid, multi-unit auction formats are best known in the financial sector for their long-time and widespread use in the sale of government securities. For example, a survey of OECD countries in 1992 found that Australia, Canada, Denmark, France, Germany, Italy, Japan, New

Zealand, the United Kingdom and, of course, the United States then used sealed-bid auctions for selling at least some of their debt. The pay-as-bid auction was the traditional format used for US Treasury bills, as well as for government securities of most other countries. The uniform-price auction was first proposed seriously as a replacement for the pay-as-bid auction by Milton Friedman in testimony at a 1959 Congressional hearing. Wilson (1979) gave the first theoretical analysis of a uniform-price auction. In 1993 the United States began an 'experiment' of using the uniform-price auction for two- and five-year government notes and, beginning in 1998, the United States switched entirely to the uniform-price auction for all issues. Meanwhile, the multi-unit Vickrey auction was introduced and first analysed in Vickrey's 1961 paper.

The pay-as-bid auction can be correctly viewed as a multi-unit generalization of the first-price auction. However, it is quite difficult to calculate Nash equilibria of the pay-as-bid auction, unless efficient equilibria exist. *Three* symmetry assumptions together guarantee the existence of efficient equilibria. First, bidders are assumed to be symmetric, in the sense that the joint distribution governing the bidders' information is symmetric with respect to the bidders. Second, bidders regard every unit of the good as symmetric: that is, each bidder i has a constant marginal valuation for every quantity $q_i \in [0, \lambda_i]$, up to a capacity of λ_i, and a marginal valuation of zero thereafter. Third, the bidders are symmetric in their capacities: that is, $\lambda_i = \lambda$, for all bidders i. With these assumptions, the pay-as-bid auction has a solution very similar to that of the first-price auction for a single item. However, without these assumptions, it inherits an undesirable property from the single-item auction: absent symmetry, all Nash equilibria of the pay-as-bid auction will generally be inefficient (Ausubel and Cramton 2002, Theorems 3 and 4).

The uniform-price auction bears a superficial resemblance to the second-price auction of a single item, in that a high winning bid gains the benefit of a lower marginal bid. However, any similarity is indeed only superficial as, except under very restrictive assumptions, all equilibria of the uniform-price auction are inefficient. The argument is simplest in the same model of constant marginal valuations as in the previous paragraph. If the capacities of all bidders are equal (that is, if $\lambda_i = \lambda$ for all i) and if the supply is an integer multiple of λ, then there exists an efficient Bayesian-Nash equilibrium of the uniform-price auction. (For example, if there are M identical units available and if every bidder has a unit demand, then sincere bidding is a Nash equilibrium in dominant strategies.) However, if the bidders' capacities are unequal or if the supply is not an integer multiple of λ, then all equilibria of the uniform-price auction are inefficient (Ausubel and Cramton 2002, Theorems 2 and 5).

The intuition for inefficiency in the uniform-price auction can be found by taking a close look at optimal bidding strategies. Sincere bidding is weakly dominant for a *first* unit: if a bidder's first bid determines the clearing price, then the bidder wins zero units. However, the bidder's *second* bid may determine the price he pays for his first unit, providing an incentive to shade his bid. The extent of *demand reduction,* as this bid shading is known, increases in the number of units, since the number of infra-marginal units whose price may be affected increases. Further, note that the allocation rule in the auction has the effect of equating the amounts of the bidders' marginal bids. Since a large bidder will likely have shaded his marginal bid more than a small bidder, the large bidder's marginal value is probably greater than a small bidder's. Consequently, the bidders' marginal values will be unequal, contrary to efficiency.

Meanwhile, the Vickrey auction is the correct multi-unit generalization of the second-price auction. As in the pay-as-bid and uniform-price auctions, bidders simultaneously submit inverse demand functions and each bidder wins the quantity demanded at the clearing price. However, rather than paying the bid price or the clearing price for each unit won, a winning bidder pays the *opportunity cost.* If a bidder wins K units, he pays the Kth highest rejected bid of his opponents for his first unit, the $(K-1)$st highest rejected bid of his opponents for his second unit, . . . , and the highest rejected bid of his opponents for his Kth unit. The dominant strategy property of the sealed-bid second-price auction generalizes

because a bidder's payment is determined solely by his opponents' bids. Consequently, given pure private values and non-increasing marginal values, sincere bidding is an efficient equilibrium in weakly dominant strategies.

Efficiency and Revenue Comparisons

Under pure private values, the dominant strategy equilibrium of the Vickrey auction attains full efficiency. It can be shown that neither the pay-as-bid nor the uniform-price auction generally attains efficiency; moreover, the efficiency ranking of these two formats is inherently ambiguous. To continue the argument of the previous subsection, it is sufficient to examine environments in which bidders have constant marginal valuations. If $F_i = F$ and $\lambda_i = \lambda$ for all bidders i, but the supply is *not* an integer multiple of λ, then the pay-as-bid auction has an efficient equilibrium while all equilibria of the uniform-price auction are inefficient. Conversely, if $\lambda_i, = \lambda$ for all bidders i and if the supply is an integer multiple of λ, but $F_i \neq F_j$ for two bidders i and j, then the uniform-price auction has an efficient equilibrium while all equilibria of the pay-as-bid auction are generally inefficient (Ausubel and Cramton 2002).

On revenues, the policy literature has generally assumed that the uniform-price auction outperforms the pay-as-bid auction; however, the argument of the previous paragraph can be extended to reverse the assumed ranking. Maskin and Riley (1989) extend Myerson's (1981) characterization of the optimal auction to multiple homogeneous goods: with symmetric bidders and constant marginal valuations, their characterization requires allocating items efficiently. Thus, as in the previous paragraph, if $F_i = F$ and $\lambda_i = \lambda$ for all bidders i, but the supply is *not* an integer multiple of λ, then the efficient equilibrium of the pay-as-bid auction outranks all equilibria of the uniform-price auction on *revenues* (as well as efficiency).

Uniform-Price Clock Auctions

The 'clock auction' – a practical design for dynamic auctions of one or more types of goods, with its origins in the 'Walrasian auctioneer' from the classical economics literature – has seen increasing use as a trading institution since 2001. A fictitious auctioneer is often presented as a device or thought experiment for understanding convergence to a general equilibrium. The Walrasian auctioneer announces a price vector, p; bidders report the quantity vectors that they wish to transact at these prices; and the auctioneer increases or decreases each component of price according as excess demand is positive or negative (*Walrasian tâtonnement*). This iterative process continues until a price vector is reached at which excess demand is zero, and trades occur only at the final price vector. In real-world applications, instead of a fictitious auctioneer serving as a metaphor for a market-clearing process, the process is taken literally; a real auctioneer announces prices and accepts bids of quantities. Applications, to date, have largely been in the electricity, natural gas, and environmental sectors.

The basic clock auction differs from the standard Sotheby's or eBay auction in that bidders do not propose prices. Rather, the auctioneer announces prices, and bidders' responses are limited to the reporting of quantities desired at the announced prices, until clearing is attained. As such, it is closest to the auction-theorist's depiction of the English auction for a single item (or the traditional Dutch auction), but generalized, so that, instead of bidders merely giving binary responses of whether they are 'in' or 'out' as prices ascend, they indicate their quantities desired.

Observe that the uniform-price clock auction is correctly viewed as a dynamic version of the sealed-bid uniform-price auction reviewed in the previous two subsections. The important difference is that, in the dynamic auction, bidders will typically receive repeated feedback as to the aggregate demand at the various prices.

As such, the clock auction may inherit the advantages that dynamic auctions have over sealed-bid auctions. First, under conditions that can be made precise, the insight from single-item auctions that feedback about other bidders' valuations would ameliorate the winner's curse and lead to more aggressive bidding carries over to the multi-unit environment. Second, clock auctions, better than sealed-bid auctions, allow bidders to maintain the privacy of their valuations for the

items being sold. Bidders never need to submit any indications of interest at any prices beyond the auction's clearing price. Third, when there are two or more types of items, auctioning them simultaneously enables bidders to submit bids based on the substitution possibilities or complementarities among the items at various price vectors. At the same time, the iterative nature of the auction economizes on the amount of information submitted: demands do not need to be submitted for all price vectors, but only for price vectors reached along the convergence path to equilibrium.

Unfortunately, the uniform-price clock auction also inherits the demand reduction and inefficiency of the sealed-bid uniform-price auction. Indeed, as a theoretical proposition, the problem of bidders optimally reducing their quantities bid well below their true demands can become substantially worse in the dynamic version of the auction. The *reductio ad absurdum* is provided by Ausubel and Schwartz (1999), who analyse a two-bidder clock auction game of complete information in which the bidders alternate in their moves. For a wide set of environments, the unique subgame perfect equilibrium has the qualitative description that, at the first move, the first player reduces his quantity to approximately half of the supply and, at the second move, the second player reduces his quantity to clear the market. Thus, the outcome is inefficient and the revenues barely exceed the starting price.

As a practical matter, demand reduction may not undermine the outcome of a uniform-price clock auction where there is substantial competition for every item being sold. However, if one or more of the bidders has considerable market power, it may become important to use an auction format which avoids creating incentives for demand reduction.

Efficient Clock Auctions

Ausubel (2004, 2006) proposes an alternative clock auction design, which utilizes the same general structure as the uniform-price clock auction, but adopts a different payment rule that eliminates the incentives for demand reduction. In essence, the design provides a dynamic version of the (multi-unit) Vickrey auction, and thereby inherits its incentives for truth-telling.

The Ausubel auction is easiest described for a homogeneous good. After each set of bidder reports, the auctioneer determines whether any bidder has 'clinched' any of the units offered (that is, whether any bidder is mathematically guaranteed to win one or more units). For example, in an auction with a supply of 5 units, and three bidders demanding 3, 2 and 2 units, respectively, the first bidder has clinched 1 unit, as his opponents' total demand of 4 is less than the supply of 5. Rather than awarding units only at a final uniform price, the auction awards units at the current price whenever they are newly clinched.

If this alternative clock auction is represented as a static auction, it collapses to the Vickrey auction in the same sense that an English auction collapses to the sealed-bid second-price auction. Consequently, it can be proven that sincere bidding is an equilibrium and, in a suitable discrete specification of the game under incomplete information, sincere bidding is the unique outcome of iterated elimination of weakly dominated strategies. Thus, unlike the uniform-price clock auction, there is no incentive for demand reduction.

Auctions of Heterogeneous Goods

In many significant applications, the multiple items offered within an auction are each unique, so it is not adequate for bidders merely to indicate the quantities that they desire. For example, an FCC spectrum auction might include a New York licence, a Washington licence and a Los Angeles licence. Moreover, there might be synergies in owning various combinations: for example, a New York and a Washington licence together might be worth more together than the sum of their values separately. Such environments pose particular challenges for auction theory.

Simultaneous Ascending Auctions

The simultaneous ascending auction, proposed in comments to the FCC by Paul Milgrom, Robert Wilson and Preston McAfee, has been used in auctions on six continents allocating more than $100 billion worth of spectrum licenses. Some of the best known applications of the simultaneous

ascending auction include: the Nationwide Narrowband Auction (July 1994), the first use of the simultaneous ascending auction; the PCS A/B Auction (December 1994–March 1995), the first large-scale auction of mobile telephone licences, which raised $7 billion; the United Kingdom UMTS Auction (March–April 2000), which raised 22.5 billion British pounds; and the German UMTS Auction (July–August 2000), which raised 50 billion euro.

In the simultaneous ascending auction, multiple items are put up for sale at the same time and the auction concludes simultaneously for all of the items. As such, it is a modern version of the 'silent auction' that is frequently used in fundraisers by charitable institutions. Bidders submit bids in a sequence of rounds. Each bid comprises a single item and an associated price, which must exceed the standing high bid by at least a minimum bid increment. After each round, the new standing high bids for each item are determined. The auction concludes after a round passes in which no new bids are submitted, and the standing high bids are then deemed to be winning bids. Payments equal the amounts of the winning bids.

The critical innovation in the simultaneous ascending auction is the inclusion of *activity rules* into the auction design. Activity rules are bidding constraints that limit a bidder's bidding activity in the current round based on his past bidding activity (that is, his standing high bids and new bids). Without activity rules, bidders would tend to wait as 'snakes in the grass' until nearly the end of the auction before placing their serious bids, thwarting any price discovery (the main reason for conducting a dynamic auction in the first place). Conversely, activity rules have the effect of forcing bidders to place meaningful bids in early rounds of the auction and thereby to reveal information to their opponents.

Walrasian Equilibria as Outcomes of Simultaneous Ascending Auctions

A Walrasian equilibrium – consisting of prices for the various items and an allocation of the items to the bidders such that each item with a non-zero price is assigned to exactly one bidder and such that each bidder prefers his assigned allocation to any alternative bundle at the given prices – is a plausible outcome for the simultaneous ascending auction. On the assumption that a Walrasian equilibrium was reached, no bidder would have any incentive to attempt to upset the allocation, even if he believed he could obtain additional items without further increasing their prices.

Thus, it becomes interesting to identify the conditions needed for existence of Walrasian equilibria with discrete items.

Kelso and Crawford (1982) show that the substitutes condition is sufficient for the existence of Walrasian equilibrium. 'Substitutes' literally refers to the price-theoretic condition that if the price of one item is increased while the price of every other item is held fixed, then the demand for every other item weakly increases. Moreover, the substitutes condition is 'almost necessary' for existence. Suppose that the set of possible bidder preferences includes all valuation functions satisfying the substitutes condition, but also includes at least one valuation function violating the substitutes condition. Then if there are at least two bidders, there exists a profile of valuation functions such that no Walrasian equilibrium exists (Gul and Stacchetti 1999; Milgrom 2000).

The reader should avoid losing sight of the fact that, just because a Walrasian equilibrium exists for a discrete environment, it does not necessarily follow that the simultaneous ascending auction will terminate at a Walrasian equilibrium. The strongest statement that can be made is that, if bidders bid 'straightforwardly' (that is, if they demand naively the bundle of items that maximizes their utility, while ignoring strategic considerations), then a Walrasian equilibrium will be reached. However, observe that, even with homogeneous goods, consumers with weakly diminishing marginal valuations satisfy the substitutes condition. Nonetheless, the uniform-price auction is susceptible to demand reduction – meaning that bidders are likely to reduce their demands and thereby end the auction before reaching a Walrasian equilibrium. Indeed, we know from the Fundamental Theorem of Welfare Economics that the Walrasian equilibrium is

efficient, so that any conclusion of inefficiency in a uniform-price auction implies that the outcome must be non-Walrasian.

Static Pay-as-Bid Combinatorial Auctions

Let us consider an example with two bidders, 1 and 2, and two items, A and B, where the substitutes condition is not satisfied and the existence of Walrasian equilibrium fails. Bidder 1 has a valuation of 3 for the package of A and B, but has a valuation of 0 for each item separately. (Thus, for Bidder 1, the goods are complements – not substitutes.) Bidder 2 has a valuation of 2 for item A, 2 for item B, and only 2 for the package of A and B. The efficient allocation assigns both items to Bidder 1. Consequently, any Walrasian equilibrium (if it exists) must assign both items to Bidder 1. However, to dissuade Bidder 2 from purchasing either item, the prices p_A and p_B of items A and B, respectively, must satisfy $p_A > 2$ and $p_B > 2$. Consequently, $p_A + p_B > 4$, exceeding Bidder 1's valuation for the package of two items and yielding a contradiction.

Given the argument of the previous paragraph, we should not expect the simultaneous ascending auction – or any auction format with bids for individual items – to generate the efficient allocation in this example. Bidder 1 's dilemma is often referred to as the exposure problem: a bidder may refrain from bidding more than his stand-alone valuations for each of the individual items, knowing that, if he is outbid on some of the individual items, he will remain 'exposed' as the high bidder on the remaining items. This may prevent the available synergies from being realized. Indeed, if Bidder 1 understands this example, he may be unwilling to bid any positive price for either item, since Bidder 2 is sure to win one of the items, and therefore Bidder 1 would obtain zero value from the item that he wins.

The exposure problem can be avoided by using a *combinatorial auction.* The rules are modified to permit bidders to place *package bids*, each comprising a *set* of items and a price. For example, the bid $(\{A, B\}, p)$ is interpreted as an all-or-nothing offer in the amount of p for the package of A and B – with no requirement that the bidder is willing to accept a part of the package for a part of the price. The allocation is determined by a combination of compatible bids that maximizes the seller's revenues. In this example, Bidder 2 is unwilling to bid any more than 2 for any combination of items, while Bidder 1 is able to exceed 2 for {A, B}. Consequently, the solution has Bidder 1 receiving both items, the efficient allocation.

To the extent that bidders value some of the items in the auction as substitutes, then it may be important for any two bids by the same bidder to be treated as *mutually exclusive.* For example, Bidder 2 in the above example may have been willing to bid 1.5 for item A and 1.5 for item B – but *not* if there was a significant risk that both bids would be accepted. This difficulty is avoided if the auction rules permit at most one of his bids to be accepted. (Such mutually exclusive bids are sometimes referred to as 'XOR' bids.) Observe that a rule of mutual exclusivity is *fully expressive* in the sense that it enables the bidder to express any arbitrary preferences. For example, if Bidder 2 in the above example wished to allow both of his bids to be accepted, he could effectively opt out of the mutual exclusivity by submitting a third bid comprising the package {A, B} at a price of 3.

In a static pay-as-bid combinatorial auction, each bidder simultaneously and independently submits a collection of package bids. The auctioneer then solves the *winner determination problem*: find a combination of bids (at most one from each bidder) that maximizes the seller's revenues subject to the constraint that each item can be allocated to at most one bidder. The submitter of each bid selected in the winner determination problem wins the items specified in the bid and pays the amount of the bid.

Rassenti et al. (1982) are credited with the first experimental study of combinatorial auctions. They studied a static combinatorial auction treating the problem of allocating airport time slots, a natural application given that landing and takeoff slots are strong complements. Bernheim and Whinston (1986) provided an important characterization of equilibria of static pay-as-bid combinatorial auctions under complete information.

The Vickrey–Clarke–Groves (VCG) Mechanism

Just as the payment rule of a pay-as-bid auction for a single item or for homogeneous goods can be modified to be 'second-price', an analogous modification can be done in the case of a combinatorial auction for heterogeneous goods. This generalization is due to Clarke (1971) and Groves (1973). Let N be an arbitrary finite set of items and let L be the set of bidders. In the *Vickrey–Clarke–Groves (VCG) mechanism,* each bidder $\ell \in L$ submits $2^{|N|}$ package bids, for all subsets of set N. After the bids are submitted, the auctioneer finds a solution, $(x_\ell)_{\ell \in L}$, to the winner determination problem. While bidder ℓ is allocated the subset $x_\ell \subset N$, he does not pay his bid $b_\ell(x_\ell)$. Rather, his payment $y_\ell \in \mathbb{R}$ is calculated so that $b_\ell(x_\ell) - y_\ell = R^*(L) - R^*(L/\ell)$, where $R^*(L)$ denotes the maximized revenue of the winner determination problem with bidder ℓ present and $R^*(L/\ell)$ denotes the maximized revenue of the winner determination problem with bidder ℓ absent. With sincere bidding, each bid $b_\ell(x_\ell)$ corresponds to the bidder's valuation $v_\ell(x_\ell)$, and $R^*(L)$ corresponds to the (maximized) social surplus. Thus, bidder ℓ is allowed a payoff equaling the *incremental surplus* that he brings to the auction. As in the Vickrey auction for homogeneous goods, a bidder's payment thus equals the opportunity cost of assigning the items to the bidder.

Applied to a setting with a single item, observe that the VCG mechanism reduces to the sealed-bid second-price auction. Applied to a setting of homogeneous goods and non-increasing marginal valuations, the VCG mechanism reduces to the (multi-unit) Vickrey auction. By the same reasoning as before, the dominance properties of these special cases extend to the setting with heterogeneous items: if bidders have pure private values, sincere bidding is a weakly dominant strategy for every bidder, yielding an efficient allocation.

Dynamic Combinatorial Auctions

In auctions for a single item, we have seen that a close relationship exists between a dynamic procedure with a pay-as-bid payment rule (that is, the English auction) and a static procedure with a second price rule (that is, the sealed-bid second-price auction). Furthermore, for homogeneous goods with non-increasing marginal values, an analogous relationship holds between the dynamic Ausubel auction and the static Vickrey auction. An important question for heterogeneous goods is the extent to which outcomes of a dynamic combinatorial auction with a pay-as-bid rule map to the static VCG mechanism.

Banks et al. (1989) conducted an early and influential study of dynamic combinatorial auctions. They defined several alternative sets of rules for the auction, developing some theoretical results and conducting an experimental study. Other important contributions have included Parkes and Ungar (2000), who independently provided a formulation of the ascending proxy auction described below, and Kwasnica et al. (2005).

Ausubel and Milgrom (2002) give two formulations of a combinatorial auction and use them to provide a partial answer to the relationship between dynamic combinatorial auctions and the VCG mechanism:

- *Ascending package auction.* Bidders submit package bids in a sequence of bidding rounds. Each new bid must exceed the bidder's prior bids for the same package by at least a minimum bid increment. After each round, the winner determination problem is solved, on all past and present bids, to determine a provisional allocation and provisional payments. The auction concludes after a round in which no new bids are submitted.
- *Ascending proxy auction.* Each bidder enters his valuations for the various packages into a *proxy bidder.* The proxy bidders then bid on behalf of the bidders in an ascending package auction in which the minimum bid increment is taken arbitrarily close to zero.

The second formulation may be viewed both as a new auction format which greatly speeds the progress of the auction, as well as a modelling device for obtaining results about the first formulation. While the first formulation is an extremely complicated dynamic game, efficiency results and

a partial equilibrium characterization are available for the second formulation.

A bidder ℓ in the ascending proxy auction is said to bid *sincerely* if he submits his true valuation, $v_\ell(S)$, for every package $S \subset N$; and he is said to bid *semisincerely* if he submits his true valuation less a positive constant, $v_\ell(S) - c$, where the same constant c is used for all packages S with valuations of at least c. The following results refer to the coalitional form game (with transferable utility) corresponding to the package economy: the value of any coalition that includes the seller is the total value associated with an efficient allocation among the buyers in the coalition; and the value of any coalition without the seller equals zero. The *core* is defined as the set of all payoff allocations that are feasible and upon which no coalition of players can improve.

Ausubel and Milgrom (2002) establish that the payoff allocation from the ascending proxy auction, given any reported preferences, is an element of the core (relative to the reported preferences). Furthermore, for any payoff vector π that is a bidder-Pareto-optimal point in the core, there exists a Nash equilibrium of the ascending proxy auction with associated payoff vector π. Conversely, for any Nash equilibrium in semi-sincere strategies at which losing bidders bid sincerely, the associated payoff vector is a bidder-Pareto-optimal point in the core.

Furthermore, the set of all economic environments essentially dichotomizes into two cases. First, if all bidders' preferences satisfy the substitutes condition, then a single point in the core dominates all other points in the core for every bidder, and it equals the payoff vector from the Vickrey–Clarke–Groves mechanism. Thus, in this first case, the outcome of the ascending proxy auction coincides with the outcome of the VCG mechanism. Second, if at least one bidder's preferences violate the substitutes condition, then there exists an additive preference profile for the remaining bidders such that there is more than one bidder-Pareto-optimal point in the core. In this second case, the VCG payoff vector is *not* an element of the core; and the low revenues of the VCG mechanism may become problematic.

Conclusion

The proportion of goods and services transacted by auction processes has dramatically increased in recent years and is likely to increase further, making the understanding of auctions and the improvement of their designs increasingly important. At the same time, auctions will remain one of the most useful test beds for game theory, since the rules of the game are better defined than in most other markets. Consequently, auction theory will almost certainly continue to be a central area of study in economics.

See Also

- ▶ Auctions (Applications)
- ▶ Auctions (Empirics)
- ▶ Auctions (Experiments)
- ▶ Incentive Compatibility
- ▶ Mechanism Design
- ▶ Vickrey, William Spencer (1914–1996)

Bibliography

Ausubel, L.M. 2004. An efficient ascending-bid auction for multiple objects. *American Economic Review* 94: 1452–1475.

Ausubel, L.M. 2006. An efficient dynamic auction for heterogeneous commodities. *American Economic Review* 96: 602–629.

Ausubel, L.M., and P. Cramton. 2002. Demand reduction and inefficiency in multiunit auctions. Working Paper 96–07, Department of Economics, University of Maryland.

Ausubel, L.M., and P.R. Milgrom. 2002. Ascending auctions with package bidding. *Frontiers of Theoretical Economics* 1(1), Article 1.

Ausubel, L.M., and J. Schwartz. 1999. *The ascending auction paradox*. Mimeo: University of Maryland.

Banks, J.S., J.O. Ledyard, and D.P. Porter. 1989. Allocating uncertain and unresponsive resources: An experimental approach. *RAND Journal of Economics* 20: 1–25.

Bernheim, B.D., and M. Whinston. 1986. Menu auctions, resource allocation and economic influence. *Quarterly Journal of Economics* 101: 1–31.

Clarke, E.H. 1971. Multipart pricing of public goods. *Public Choice* 11: 17–33.

Cramton, P., Y. Shoham, and R. Steinberg. 2006. *Combinatorial auctions*. Cambridge, MA: MIT Press.

Crémer, J., and R.P. McLean. 1985. Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica* 53: 345–361.
Groves, T. 1973. Incentives in teams. *Econometrica* 41: 617–631.
Gul, F., and E. Stacchetti. 1999. Walrasian equilibrium with gross substitutes. *Journal of Economic Theory* 87: 95–124.
Harris, M., and A. Raviv. 1981. Allocation mechanisms and the design of auctions. *Econometrica* 49: 1477–1499.
Kelso, A.S. Jr., and V.P. Crawford. 1982. Job matching, coalition formation, and gross substitutes. *Econometrica* 50: 1483–1504.
Klemperer, P. 2000. *The economic theory of auctions*. Cheltenham: Edward Elgar.
Krishna, V. 2002. *Auction theory.* San Diego: Academic Press.
Kwasnica, A.M., J.O. Ledyard, D. Porter, and C. DeMartini. 2005. A new and improved design for multiobject iterative auctions. *Management Science* 51: 419–434.
Maskin, E., and J. Riley. 1989. Optimal multi-unit auctions. In *The economics of missing markets, information, and games*, ed. F. Hahn. Oxford: Oxford Univ. Press.
Maskin, E., and J. Riley. 2003. Uniqueness of equilibrium in sealed high-bid auctions. *Games and Economic Behavior* 45: 395–409.
McAfee, R.P., and J. McMillan. 1987. Auctions and bidding. *Journal of Economic Literature* 25: 699–738.
Milgrom, P. 2000. Putting auction theory to work: The simultaneous ascending auction. *Journal of Political Economy* 108: 245–272.
Milgrom, P. 2004. *Putting auction theory to work*. Cambridge: Cambridge University Press.
Milgrom, P.R., and R.J. Weber. 1982. A theory of auctions and competitive bidding. *Econometrica* 50: 1089–1122.
Myerson, R.B. 1981. Optimal auction design. *Mathematics of Operations Research* 6: 58–73.
Parkes, D.C., and L.H. Ungar. 2000. Iterative combinatorial auctions: theory and practice. *Proceedings of the 17th National Conference on* Artificial Intelligence, 74–81.
Rassenti, S.J., V.L. Smith, and R.L. Bulfin. 1982. A combinatorial auction mechanism for airport time slot allocation. *Bell Journal of Economics* 13: 402–417.
Riley, J.G., and W.F. Samuelson. 1981. Optimal auctions. *American Economic Review* 71: 381–392.
Vickrey, W. 1961. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance* 16: 8–37.
Wilson, R. 1979. Auctions of shares. *Quarterly Journal of Economics* 93: 675–689.
Wilson, R. 1992. In *Strategic analysis of auctions. In handbook of game theory*, 1st ed., ed. R. Aumann and S. Hart. Amsterdam: North-Holland.

Aumann, Robert J. (Born 1930)

Abraham Neyman

Abstract

Robert Aumann has played an essential and indispensable role in shaping game theory and much of economic theory. He promotes a unified view of the very wide domain of rational behaviour, a domain that encompasses areas of many apparently disparate disciplines, like economics, political science, biology, psychology, mathematics, philosophy, computer science, law and statistics. His contributions have had a most profound impact on the social sciences

Keywords

Aumann, R. J; Bounded rationality,; Coalitions; Common knowledge; Continuum of traders; Contract curve; Convexity; Core; Core equivalence; Correlated equilibrium; Folk theorem; Game theory; Non-transferable utility; Perfect competition; Rational behaviour; Repeated games with incomplete information; Shapley value; Strong equilibrium; Supergames; Transferable utility

JEL Classification

B31

Robert J. Aumann, Professor Emeritus of Mathematics at the Hebrew University of Jerusalem, and member of the interdisciplinary Center for Rationality there, shares (with Thomas C. Schelling) the 2005 Nobel Prize in Economics (Aumann and Schelling 2005).

Aumann was born in Frankfurt, Germany, in 1930, and moved to New York with his family in 1938. In 1955 he completed his Ph.D. in mathematics at MIT under the supervision of George Whitehead. His thesis, in knot theory, was published in the *Annals of Mathematics* (Aumann 1956).

In 1955, Aumann joined the Princeton University group that worked on industrial and military applications, where he realized the importance and relevance of game theory, then in its infancy. In 1956 Aumann joined the Institute of Mathematics at the Hebrew University.

Since the mid-1950s, Aumann has played an essential and indispensable role in shaping game theory, and much of economic theory, to become the great success it is today. He promotes a unified view of the very wide domain of rational behaviour, a domain that encompasses areas of many apparently disparate disciplines, like economics, political science, biology, psychology, mathematics, philosophy, computer science, law, and statistics. Aumann's research is characterized by an unusual combination of breadth and depth. His scientific contributions are path-breaking, innovative, comprehensive and rigorous, ranging from the discovery and formalization of the basic concepts and principles, through the development of the appropriate tools and methods for their study, to their application in the analysis of various specific issues. Some of his contributions require very deep and complex technical analysis; others are (as he says at times) 'embarrassingly trivial' mathematically, but very profound conceptually. He has influenced and shaped the field through his pioneering work. There is hardly an area of game theory today where his footprint is not readily apparent. Most of Aumann's research is intimately connected to central issues in economic theory; on the one hand, these issues provided the motivation and impetus for his work; on the other, his results produced novel insights and understandings in economics. No less important than his own pioneering work is Aumann's indirect impact through his many students, collaborators and colleagues. He inspired them, excited them with his vision, and led them to further important results.

Here we must confine ourselves to brief commentary touching on only a small part of his output. It is important to note that the scope of each description is not indicative of the importance of the contribution. Further and more detailed accounts of Aumann's contributions may be found in Hart and Neyman (1995).

We start with Aumann's study of long-term interactions, which had a most profound impact on the social sciences. The mathematical model enabling a formal analysis is a *supergame* G^*, consisting of an infinite repetition of a given one-stage game G. (A game G in strategic form consists of a set of players N, pure strategy sets Ai for each player i, and payoff functions gi, which describe the payoff to player i as a function of the strategy profiles $a \in A := Xi \in NAi$.) A pure strategy in G^* assigns a pure strategy in G to each period/stage, as a function of the history of play up to that stage. A profile of supergame strategies, one for each player, defines the play, or sequence of stage actions. The payoff associated with a play of the supergame is essentially an average of the stage payoffs.

In 1959 Aumann defined the notion of a *strong equilibrium* – a strategy profile where no group of players can gain by unilaterally changing their strategies – and characterized the strong equilibrium outcomes of the supergame by showing that it coincides with the so-called *β-core* of G. When Aumann's 1959 methodology is applied to *Nash equilibrium* – a strategy profile where no single player can gain by unilaterally changing his strategy – the result is essentially the so-called *folk theorem* for supergames: the set of Nash equilibria of the supergame G^* coincides with the set of feasible and individual rational payoffs in the one-stage game. In 1976, Aumann and Shapley (and Rubinstein 1976, in independent work) proved that the equilibrium payoffs and the perfect equilibrium payoffs of the supergame G^* coincide.

Supergames are repeated games of complete information; it is assumed that all players know precisely the one-shot game that is being repeatedly played.

> The theory of repeated games of complete information is concerned with the evolution of fundamental patterns of interaction between people (or for that matter, animals; the problems it attacks are similar to those of social biology). Its aim is to account for phenomena such as cooperation, altruism, revenge, threats (self-destructive or otherwise), etc. – phenomena which may at first seem irrational – in terms of the usual 'selfish' utility-maximizing paradigm of game theory and neoclassical economics. (Aumann 1981, p. 11)

The model of repeated games with incomplete information, introduced in 1966 by Aumann and Maschler (Aumann and Maschler 1995), analyses long-term interactions in which some or all of the players do not know which stage game G is being played. The game $G = Gk$ depends on a parameter k; at the start of the game a commonly known lottery $q(k)$ with outcomes in a product set $S = \times iSi$ is performed and player i is informed of the i-th coordinate of the outcome. The repetition enables players to infer and learn information about the other players from their behaviour, and therefore there is

> a subtle interplay of concealing and revealing information: concealing, to prevent the other players from using the information to your disadvantage; revealing, to use the information yourself, and to permit the other players to use it to your advantage. (Aumann 1985, pp. 46–47)

> The stress here is on the strategic use of information – when and how to reveal and when and how to conceal, when to believe revealed information and when not, etc. (Aumann 1981, p. 23)

This problem of the optimal use of information is solved in an explicit and elegant way in Aumann and Maschler (1995).

Another substantial line of contributions of Aumann is the introduction and study of the continuum idea in game theory and economic theory.

A perfectly competitive economic model is meant to describe a situation in which there are many participants, and the influence of each one individually is negligible. The state of the economy is thus insensitive to the actions of any single agent; only the aggregate behaviour matters. For instance, in a pure exchange economy in which the initial endowment of each trader is very small relative to the whole, the quantities of goods traded by any one agent cannot essentially affect the total supply and demand.

The first question is: What is the correct way of modelling perfect competition? Aumann introduced the model of economies with a continuum of participants, as the appropriate model where each individual is indeed insignificant:

> Indeed, the influence of an individual participant on the economy cannot be mathematically negligible, as long as there are only finitely many participants. Thus a mathematical model appropriate to the intuitive notion of perfect competition must contain infinitely many participants. We submit that the most natural model for this purpose contains a continuum of participants, similar to the continuum of points on a line or the continuum of particles in a fluid. (Aumann 1964, p. 39)

The introduction of the 'continuum' idea in economic theory has been indispensable to the advancement of this discipline. In the same way as in most of the natural sciences, it enables a precise and rigorous analysis, which otherwise would have been very hard or even impossible. Specifically,

> the continuum can be considered an approximation to the 'true' situation in which there is a large but finite number of particles (or traders, or strategies, or possible prices). The purpose of adopting the continuous approximation is to make available the powerful and elegant methods of the branch of mathematics called 'analysis,' in a situation where treatment by finite methods would be much more difficult or even hopeless (think of trying to do fluid mechanics by solving n-body problems for large n. (Aumann 1964, p. 41)

Once the basic model is specified, the next question is: What does perfect competition lead to? The classical economic approach is that there are prices for all goods, which every agent takes as given (he is, after all, insignificant, so his decision cannot affect the prices). In order for the economy to be in a stable situation the prices must be such that the total demand equals the total supply. This is the Walrasian competitive equilibrium. That it exists and is well defined in markets with a continuum of traders was shown by Aumann in 1966; moreover, unlike in finite markets, no convexity assumptions were required.

Another approach considers the possible trades that groups of agents – called coalitions – can make among themselves, in such a way that they all benefit. This leads to the core, a game-theoretic concept that generalizes Edgeworth's famous 'contract curve': the core consists of all those allocations that no coalition can improve upon. These are clearly different concepts:

> The definition of competitive equilibrium assumes that the traders allow market pressures to determine prices and that they then trade in accordance with these prices, whereas that of core ignores the price

> mechanism and involves only direct trading between the participants. (Aumann 1964, p. 40)

Aumann (1964) showed that the core and the set of competitive allocations coincide in markets with a continuum of traders. By introducing the model of the continuum that expresses precisely the idea of perfect competition, he succeeded in making precise also this equivalence (originally suggested by Edgeworth 1881, and proved in various other models – Shubik 1959; Debreu and Scarf 1963), which has since become one of the basic tenets of economic theory.

Aumann then turned to the study of other concepts in the context of perfectly competitive markets. A traditional idea in economics is that of 'marginal worth' or 'marginal contribution'. This idea is embodied in the concept of value due to Lloyd Shapley (1953). It may be interpreted as follows:

> The Shapley value is an a priori measure of a game's utility to its players; it measures what each player can expect to obtain, 'on the average,' by playing the game. Other concepts of cooperative game theory... predict outcomes (or sets of outcomes) that are in themselves stable, that cannot be successfully challenged or upset... The Shapley value... can be considered a mean, which takes into account the various power relationships and possible outcomes. (Aumann 1978, p. 995)

While the definition of competitive equilibrium or core generalizes in a straightforward manner to the continuum of players case, this is not so in the case of value. This led to a most prolific collaboration between Aumann and Shapley, starting in the late 1960s and culminating in 1974 with the publication of their book *Values of Non-Atomic Games*. They addressed deep problems, both conceptual – how to define the correct notions – and technical, and solved them masterfully. In consequence, most important and beautiful insights were obtained. One example is the 'diagonal principle', stating that in games with many players one need consider only coalitions whose composition constitutes a good sample of the grand coalition of all participants. It is important to note that, unlike the core (or the competitive equilibrium), the value solution is applicable in almost every interactive set-up. For instance, political contexts usually lead to situations where the core is empty, whereas the value is well defined and yields most significant insights.

Returning to perfectly competitive economies, in 1975 Aumann obtained another equivalence result, this time between the competitive allocations and the value allocations – on the assumption that the market is 'sufficiently smooth'. (Again, the continuum of traders model allows Aumann to obtain a precise and general result; the first such result, in transferable utility markets only, is due to Shapley 1964.) This is perhaps even more surprising than the core equivalence, since the concept of value does not capture, by its definition, considerations of stability and equilibrium.

This equivalence is indeed striking. In Aumann's view:

> Perhaps the most remarkable single phenomenon in game and economic theory is the relationship between the price equilibria of a competitive market economy, and all but one of the major solution concepts for the corresponding game.... Intuitively, the equivalence principle says that the institution of market prices arises naturally from the basic forces at work in a [perfectly competitive] market, (almost) no matter what we assume about the way in which these forces work. (From game theory)

This nicely exemplifies Aumann's view on the universality of the game theoretic approach:

> The more conventional approaches take institutions as given, and ask where they lead. The game-theoretic approach asks how the institutions came about, what led to them? Thus general equilibrium theory takes the idea of market prices for granted; it concerns itself with their existence and properties, calculating them, and so on. Game Theory asks, why are there market prices? How did they come about? (From game theory)

The fundamental insights and understandings obtained in the analysis of perfect competition enabled and facilitated the study of basic economic issues that go beyond perfect competition. We mention a few where Aumann's contributions and influence are most noticeable: monopolistic and oligopolistic competition, modelled by a continuum of traders together with one or more large participants (Shubik 1959); public economics – models of taxation based on the interweaving of the economic activities with a political process, such as voting (Aumann and Kurz 1977a, b;

Aumann et al. 1977, 1983, 1987); fixed-price models (Aumann and Drèze 1986).

Another fundamental contribution of Aumann is 'Agreeing to Disagree' (1976): it formalizes the notion of *common knowledge* and shows (the somewhat unintuitive result) that, if two agents start with the same prior beliefs and their posterior beliefs (about a specific event), which are based on different private information, are common knowledge, then these posterior beliefs coincide. This paper had a major impact; it led to the development of the area known as *interactive epistemology* and has found many applications in different disciplines like economics and computer science.

Other fundamental contributions include the introduction and study of *correlated equilibrium*, the study of *bounded rationality*, and many important contributions to cooperative game theory: extending the theory of *transferable utility* (TU) games to general *nontransferable utility* (NTU) games, formulating a simple set of axioms that characterize the NTU-value (introduced in Shapley 1969) and the 'Game-Theoretic Analysis of a Bankruptcy Problem from the Talmud' (Aumann and Maschler 1985).

Aumann has been a Member of the US National Academy of Sciences since 1985, a Member of the Israel Academy of Sciences and Humanities since 1989, a Foreign Honorary Member of the American Academy of Arts and Sciences since 1974, and a corresponding fellow of the British Academy since 1995. He received the Harvey Prize in Science and Technology in 1983, the Israel Prize in Economics in 1994, the Lanchester Prize in Operations Research in 1995, the Nemmers Prize in Economics in 1998, the EMET prize in Economics in 2002, the von Neumann prize in Operations Research in 2005, and the Nobel Memorial Prize in Economic Sciences in 2005. He was awarded honorary doctorates by the University of Bonn in 1988, by the Université Catholique de Louvain in 1989, and by the University of Chicago in 1992.

See Also

- ▶ Convexity
- ▶ Game Theory

Selected Works

1956. Asphericity of alternating knots. *Annals of Mathematics* 64, 374–392.

1959. Acceptable points in general cooperative *n*-person games. In *Contributions to the theory of games*, vol. 4 (AM-40), ed. A.W. Tucker and R.D. Luce. Princeton: Princeton University Press.

1964. Markets with a continuum of traders. *Econometrica* 32, 39–50.

1966. Existence of competitive equilibria in markets with a continuum of traders. *Econometrica* 34, 1–17.

1974. (With L.S. Shapley.) *Values of Non-Atomic Games*. Princeton: Princeton University Press.

1975. Values of markets with a continuum of traders. *Econometrica* 43, 611–646.

1976. Agreeing to disagree. *Annals of Statistics* 4, 1236–1239.

1976. (With L.S. Shapley.) Long-term competition: A game-theoretic analysis.
Mimeo, Hebrew University;repr. In *Essays in game theory in honor of Michael Maschler*, ed. N. Megiddo. Berlin: Springer, 1994.

1977. (With R.J. Gardner and R.W. Rosenthal.) Core and value for a public goods economy: An example. *Journal of Economic Theory* 15, 363–365.

1977a. (With M. Kurz.) Power and taxes. *Econometrica* 45, 1137–1161.

1977b. (With M. Kurz.) Power and taxes in a multi-commodity economy. *Israel Journal of Mathematics* 27, 185–234.

1978. Recent developments in the theory of the Shapley value. In *Proceedings of the international congress of mathematicians*. Helsinki: Academia Scientiarum Fennica, 1980.

1981. Survey of repeated games. In *Essays in game theory and mathematical economics in honor of Oskar Morgenstern*, vol. 4. Mannheim: Bibliographisches Institut.

1983. (With M. Kurz and A. Neyman.) Voting for public goods. *Review of Economic Studies* 50, 677–694.

1985. What is game theory trying to accomplish? In *Frontiers of economics*, ed. K.J. Arrow and S. Honkapohja. Oxford: Basil Blackwell.

1985. (With M. Maschler.) Game-theoretic analysis of a bankruptcy problem from the Talmud. *Journal of Economic Theory* 36, 195–213.
1986. (With J.H. Drèze.) Values of markets with satiation or fixed prices. *Econometrica* 54, 1271–318.
1987. (With M. Kurz and A. Neyman.) Power and public goods. *Journal of Economic Theory* 42, 108–127.
1995. (With M. Maschler and the collaboration of R.E. Stearns.) *Repeated games of incomplete information*. Cambridge, MA: MIT Press.
2000. *Collected papers*, vols 1 and 2. Cambridge, MA: MIT Press.
2005. (With T. Schelling.) *Robert Aumann's and Thomas Schelling's contributions to game theory: Analyses of conflict and cooperation*. Stockholm: Royal Swedish Academy of Sciences. Online. Available at http://nobelprize.org/economics/laureates/2005/ecoadv05.pdf. Accessed 5 June 2007.

Bibliography

Debreu, G., and H. Scarf. 1963. A limit theorem on the core of an economy. *International Economic Review* 4: 236–246.
Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.
Hart, S., and A. Neyman. 1995. Introduction. In *Games and economic theory, selected contributions in honor of Robert J. Aumann*, ed. S. Hart and A. Neyman. Ann Arbor: University of Michigan Press. Online. Available at http://ratio.huji.ac.il/dp/neyman/bookintroduction95.pdf. Accessed 5 June 2007.
Neyman, A. 2006. Aumann awarded Nobel prize. *Notices of the AMS* 53: 44–46.
Rubinstein, A. 1976. *Equilibrium in supergames*. RM-26. Hebrew University of Jerusalem.
Shapley, L.S. 1953. A value for *n*-person games. In *Contributions to the theory of games*, vol. 2, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press.
Shapley, L.S. 1964. *Values of large games VII: A general exchange economy with money*, RM-4248. Santa Monica: RAND Corporation.
Shapley, L.S. 1969. Utility comparison and the theory of games. In *La Décision*. Paris: CNRS.
Shubik, M. 1959. Edgeworth market games. In *Contributions to the theory of games*, vol. 4, ed. A.W. Tucker and R.D. Luce. Princeton: Princeton University Press.

Aupetit, Albert (1876–1943)

Roger Dehem

Keywords
Aupetit, A.; Bimetallism; Composite money standards; Exchange rate determination; Index numbers; Quantity theory of money; Walras, L.

JEL Classifications
B31

Aupetit was born in Sancerre (Cher). His two doctoral theses at the Faculté de Droit were respectively entitled *Théorie générale de la monnaie* (1901) and *Les accidents du travail dans l'agriculture.* Having twice failed the *concours d'agrégation,* the narrow gateway to a professorship at the Faculté de Droit, he entered the research department at the Banque de France, where he served as secretary-general from 1920 to 1926. He then entered private business. In 1936 he was elected a member of the Institut de France. His teaching was restricted to the Ecole Pratique des Hautes Etudes (1910–14) and to the Ecole des Sciences Politiques, from 1921 on.

Considered by Walras as his first disciple in France, Aupetit can best be judged by the master himself: 'He is in agreement with my social economics as well as with my pure and applied economics. He is the best and most brilliant disciple and successor I may wish to have' (Jaffé 1965, p. 353). Aupetit's *Essai sur la théorie générale de la monnaie* is a faithful though simpler and more precise reformulation of Walras' general equilibrium and monetary theories. The postulates sustaining the quantity theory are made remarkably explicit. Questions of composite monetary standards, bimetallism, exchange rate determination and index numbers are also thoroughly discussed.

Selected Works

1901. *Essai sur la théorie générale de la monnaie.* Paris: Guillaumin. A truncated version of this book was published by Marcel Rivière, Paris, in 1957.
1905. L'oeuvre économique de Cournot. *Revue de métaphysique et de morale* 13: 377–393.

Bibliography

Jaffé, W. 1965. *Correspondence of Léon Walras and related papers*. Vol. 3. Amsterdam: North-Holland.

Auspitz, Rudolf (1837–1906)

Jürg Niehans

Keywords

Auspitz, R.; Austrian School; Böhm-Bawerk, E. von; Cartels; Competitive equilibrium; Consumer surplus; Disutility; Edgeworth, F. Y.; Forward markets; Gains from trade; Indifference curves; Indivisibilities; Inventories; Launhardt, C. F. W.; Lieben, R.; Marginal utility of money; Marshall, A.; Mathematical economics; Menger, C.; Multiple equilibria; Options; Pareto, V.; Partial equilibrium; Producer surplus; Reciprocal demand curves; Security markets; Speculation; Substitutes and complements; Technical progress; Walras, L.; Wicksell, J. G. K

JEL Classifications

B31

Auspitz was born on 7 July 1837 in Vienna, where he died on 8 March 1906. He grew up in a well-educated Jewish family and studied mathematics and physics but without acquiring a degree. At the age of 26, apparently with some reluctance, he became a businessman and founded one of the first sugar refineries of the Austrian empire. As a lifelong opponent of cartels, he used to donate the extra profits he obtained from the sugar cartel to the employees' pension fund. Auspitz was also Richard Lieben's partner in the family bank, Auspitz, Lieben & Co.

A successful Liberal politician, Auspitz was a member of the Moravian Diet (1871–1900) and of the Austrian lower chamber (1873–90 and 1892–1905), where he acquired a reputation and influence as a financial expert. His first wife was Lieben's sister and a first cousin. They had two children, but the marriage was dissolved after 20 years because of the wife's insanity, whereupon Auspitz married his children's governess. He seems to have been a man of quiet energy and balanced judgement, untiring but of frail health. In some respects his life reminds one of Ricardo's.

All of Auspitz's significant scientific work was done jointly with Lieben; nothing seems to be known about their relative contributions. In 1889 appeared the *Researches on the Theory of Price,* the book that assured its authors of a place among the eminent mathematical economists. It is essentially an exhaustive partial-equilibrium analysis of price in terms of an ingenious geometrical apparatus.

The fundamental first chapter, preprinted in 1887 to fix priorities relative to Böhm-Bawerk, provides the basic tools. For every quantity of a given commodity, the 'curve of total satisfaction' indicates the maximum amount of money the buyer is willing to pay. The 'total cost curve', on the other hand, plots the minimum amount of money for which the seller (producer) is willing to supply each quantity. In modern terminology, these are indifference curves. The corresponding marginal curves, called respectively demand and supply curves, give the maximum (minimum) amount of money for which the buyer (seller) is willing to buy (sell) an additional unit.

On the assumption of a constant marginal utility of money, both parties choose the quantity in such a way that this marginal value is equal to the market place. The two marginal curves are thus equivalent to Marshall's reciprocal demand curves as applied to the exchange of one

commodity against money. Auspitz and Lieben did not know Marshall's privately printed paper of 1879, however.

Competitive equilibrium is established where the demand curve intersects the supply curve. The vertical distances between the equilibrium point and the two indifference curves then measure the gains from trade, which leads to an analysis of consumer's and producer's surplus (but without these terms).

In subsequent chapters this apparatus is applied to a wide range of microeconomic problems and cases, including substitutes and complements, indivisibilities, disutility, technical progress, inventories, security markets, forward markets and options. Among many notable pieces of analysis one finds the argument that speculation is socially beneficial if it is profitable, and a derivation of long-run curves as envelopes of short-run curves which was not surpassed until Harrod and Viner. An important final chapter extends the analysis to monopoly, monopolistic competition, excise taxes and international trade, and includes a brilliant discussion of optimal tariffs (which disturbed free-trader Pareto; see *Giornale degli Economisti,* 1892).

Four appendices present the main argument in terms of univariate differential calculus, concluding with an extension to general equilibrium. In contrast to Launhardt, who, as an engineer, loved to computed numerical results for special functional forms, Auspitz and Lieben emphasize the logic of the problem.

Auspitz and Lieben, though highly regarded by men like Edgeworth, Pareto and Fisher, never received the credit they deserved. In their local environment, in view of the Austrian School's intolerance for mathematics, they were academic outcasts. This is illustrated by Menger's critical review (*Wiener Zeitung,* 8 March 1889, quoted in Weinberger 1931) and by Auspitz's exchange with Böhm-Bawerk of 1894, which also shows Auspitz's analytical superiority. More importantly, Auspitz and Lieben, cut off from direct scholarly intercourse, were prisoners of their idiosyncrasy, never developing the knack for felicitous terminology and expository devices that in economics is so important for academic success. It also turned out that for partial analysis Cournot's price/quantity diagram is often more illuminating than the reciprocal demand curves.

Despite their gentle, scholarly personalities, Auspitz and Lieben also managed to stir up a controversy with Walras (see *Correspondence of Léon Walras and Related Papers,* ed. William Jaffé, 3 vols, Amsterdam, 1965). As early as 1887, Launhardt had warned Walras of the 'plagiarism' of those 'insolent Jewish pirates'. The preface to the *Researches,* while revealing Launhardt's diatribes as entirely unfounded, added a more substantive irritant by arguing that (1) Walras' simultaneous demand curves were not correctly constructed, in as much as the curve for one good presupposes a given price for the other, and (2) there cannot be multiple equilibria. This criticism stung Walras all the more since Edgeworth, in his presidential address of 1889, described Auspitz and Lieben as more accurate than Walras (an unwarranted observation, deleted in *Papers Relating to Political Economy*). Walras tried to mobilize Pareto and Bortkiewicz in his defence (without success) and began to polemicize against those who 'make bad theory in mathematical language'. His own reply, however (reprinted in the 4th edition of the '*Eléments*'), missed the essential point and only added to the confusion. Wicksell, as usual, got things right (*Wert, Kapital und Rente,* 1893). Auspitz and Lieben had overlooked the fact that Walras' curves, in effect, related to the demand and supply of one good in terms of the other, and the impossibility of multiple equilibria depended on the constancy of the marginal utility of money. After Auspitz's death Lieben graciously acknowledged their error (to which Walras, ungraciously, replied that the point was not important after all).

See Also

► Lieben, Richard (1842–1919)

Selected Works

1885. *Meine parlamentarische Thätigkeit während der Reichsraths-Session 1879–85.* Vienna.

1887. (With R. Lieben.) *Zur Theorie des Preises.* Leipzig: Duncker & Humblot.
1889. (With R. Lieben.) *Untersuchungen über die Theorie des Preises.* Leipzig: Duncker & Humblot. French translation by Louis Suret, Paris: M. Giard & E. Brière, 1914.
1890a. (With R. Lieben.) Reply. (To article by L. Walras in same publication.) *Revue d'économie politique* 4.
1890b. Die klassische Werttheorie und die Theorie von Grenznutzen. *Jahrbücher für Nationalökonomie und Statistik* 55.
1894. Der letzte Maasstab des Güterwertes und die mathematische Methode. *Zeitschrift für Volkswirtschaft, Socialpolitik und Verwaltung* 3.

Bibliography

Weinberger, O. 1931. Rudolf Auspitz und Richard Lieben. *Zeitschrift für die gesamte Staatswissenschaft* 91.
Weinberger, O. 1935. Rudolf Auspitz. In *Neue Oesterreichische Biographie 1815–1918*, vol. VIII. Vienna.
Winter, J. 1927. *Fünfzig Jahre eines Wiener Hauses*. Vienna: F. Jasper.

Australasia, Economics in

Selwyn Cornish

Abstract

Writing on economic subjects began in Australasia (Australia and New Zealand) within a decade or two of the commencement of European settlement. There are many examples of innovative and influential contributions to economics from these countries, but there has never been a 'school' of Australasian economics. Between the two world wars, economics in Australia experienced a golden age, when a small group of economists influenced economic policy and advanced economic thought. Since the 1940s, however, Australasian economics has been dominated by ideas and methods associated with work in the United States.

Keywords

Arndt, H. W.; Australasia, economics in; Brigden, J.B.; Butlin, N.; Capital theory; Clark, C. G.; Coghlan, T.A.; Condliffe, J.B.; Copland, D.; Corden, W. M.; Crawford, J. G.; De Lissa, A.; Dutch disease; Economic Society of Australia; Effective protection; Equation of exchange; Fisher, A.G.B.; Full employment; Garnaut, R.; General equilibrium trade models; Giblin, L. F.; Gregory thesis; Growth, models of; Harcourt, G. C.; Hearn, W. E.; Hight, J.; Inflation; International monetary economics; International trade theory; Irvine, R. F.; Jevons, W. S.; Kemp, M.; Keynes, J. M.; Marshall, A.; Measurement; Mill, J. S.; Money; Multiplier; Musgrave, A.; National income accounting; New Zealand Economic Association; Phillips, A.W.; Premiers' Plan (Australia); Protection; Salter, W.E.G.; Shann, E.; Snooks, G. D.; Swan, T. W.; Syme, D.; Tariffs; Tradable and non-tradable commodities; Unemployment; Viner, J.; Wakefield, E.G.; Walker, F. A.; Wilson, R.

JEL Classifications

B2

There has never been a 'school' of Australasian economics in the sense that English, German, Austrian, Italian, American and Swedish schools are said to have existed.

This is not to say that Australians and New Zealanders have contributed little or nothing to the history of economics. On the contrary, an economics literature commenced from the early decades of the 19th century. For the most part, economic analysis was derived from ideas originating outside the region, though imported ideas were adapted, extended and refashioned to meet peculiar Australasian conditions and circumstances. Between the two world wars, economics in Australia experienced a golden age when a remarkable group of economists exerted a profound impact on economic policy, and in the process advanced economic thought. Since the Second World War, Australasian economics has been dominated by approaches and methods that

are characteristically associated with the discipline in the United States, a phenomenon by no means unique to Australia and New Zealand.

The 19th Century

Survival was difficult and far from guaranteed for some years immediately after the establishment of European settlement in Australia in 1788. In these circumstances there was little time to write about economics. But as private activity evolved from the original penal settlements, economic issues were debated more frequently. By the 1840s, a flourishing private economy had developed around the wool export trade with Britain. The pastoral industry was land intensive, giving rise to discussion about the occupation and alienation of crown land. The growth of domestic production led to an interest in its measurement and the contributions made by different industries. The creation of private institutions, especially those catering to foreign trade, including banks and other financial institutions, wholesaling and retailing, shipping and inland transport, became subjects of interest among those who wrote and talked about economic matters. Population growth and immigration were other subjects that drew attention. With the rise of domestic and foreign trade, instability occasioned by excessive optimism and pessimism was manifested in booms and slumps; this, too, engaged the interest of writers.

E.G. Wakefield, though he never visited the antipodes, wrote in 1829 that the Australian colonies were in a barbarous condition, like that of every people scattered over a territory immense in proportion to their numbers; every man is obliged to occupy himself with questions of daily bread; there is neither leisure nor reward for investigation of abstract truth; money-getting is the universal object; taste, science, morals, manners, abstract politics are subjects of little interest unless they bear on the wool question. (Quoted in Nadel 1957, p. 36)

There is some truth in this, but, by the time Wakefield wrote, pamphlets and books by colonists on economic topics had started to appear. In 1819, for example, W.C. Wentworth published *A Statistical, Historical, and Political Description of the Colony of New South Wales and its Dependent Settlements in Van Diemen's Land*. Wentworth estimated the national income of New South Wales and Van Diemen's Land (since renamed Tasmania), and discussed processes of economic development that borrowed heavily from Adam Smith. Another early writer of some significance was the Reverend John Dunmore Lang. In 1834 he published *An Historical and Statistical Account of New South Wales* which provided a description of economic progress in the colony and an analysis of the nature and causes of the depressions of the late 1820s and the early 1840s.

William Stanley Jevons spent some years in Australia in the 1850s as assayer to the Royal Mint in Sydney. He wrote on railways and land development, and commenced a social survey of Sydney, revealing some of the promise that later was to emerge in his work in economics. Perhaps the most important writer on economics in Australia during the second half of the century was William Edward Hearn. Born in Ireland and educated at Trinity College, Dublin, an exact contemporary of Cairnes and Cliffe Leslie, Hearn in 1854 was appointed foundation Professor of Modern History, Modern Literature, Logic and Political Economy in the University of Melbourne. As an academic (he later became a Member of Parliament), Hearn published a number of books, of which the most important was *Plutology* (1863). Written as a university textbook, it was widely known in Britain and elsewhere as an outstanding summary of the state of economic knowledge. Hearn believed that the satisfaction of wants, and the efforts to meet them, constituted the chief problems of economics.

Another prominent writer of the second half of the 19th century was Sir Anthony Musgrave, Governor of South Australia and later of Queensland. His major work, *Studies in Political Economy* (1875), contained six essays critical of J.S. Mill. He claimed that Mill had failed to explore adequately the role of money as a store of value and there were deficiencies in Mill's discussion of capital. Though Musgrave's work

was often quoted, his jaundiced view of Mill's writing won him few friends among authorities overseas. David Syme, proprietor of *The Age*, a Melbourne newspaper, was yet another writer with a reputation beyond Australia. Better known for his powerful advocacy of protection, and for his writing on the disposal of crown land, Syme published as well on economic methodology and other abstract topics.

His *Outlines of an Industrial Science* (1876) seems to have been known in Europe, notably in Germany. Syme supported the application of inductive approaches to economics and criticized Mill for arguing that economics should be based on deduction. He wrote as well on economic motivation and on supply and demand analysis, criticizing as he did Mill's theory of value.

Towards the end of the 19th century a number of factors combined to encourage greater scrutiny of economic issues. One was the banking and financial crisis and collapse of economic activity in eastern Australia in the 1890s. As a consequence of the depression, debate sharpened on subjects such as the causes of fluctuations in economic activity, the role of government in moderating booms and slumps, the need for a central or government bank, unemployment and tariff policy. Another issue was the projected federation of the Australian colonies. Hitherto the six colonies of Australia had acted independently, having their own administrations, including armies and navies. Ever since the middle of the 19th century there had been calls for an Australian federation; during the 1890s several inter-colonial conventions were held to draft a federal constitution, at which economic and financial considerations, including tariffs, taxation, federal–state finance, money and banking, were debated at length.

Reflecting the heightened interest in economics for these and other reasons, an Australian Economic Association was formed in Sydney in 1887. Between March 1888 and December 1898 the Association published a monthly periodical (for a short time it was published fortnightly). Contributors to the *Australian Economist* were interested principally in the issues of the day, including unemployment, wage rates, tariff policy, recovery measures, control of banks and money, land tenure, federation, socialism, state banks, education, immigration, the role of women, democracy, bimetallism, old age pensions and industrial arbitration. Short extracts from the works of prominent economists, including Jevons, Marshall and F.A. Walker, were often included, as were articles about the work of these and other economists.

The most original of the local contributors to the *Australian Economist* was Alfred De Lissa, whose work sometimes is heralded as a forerunner of the multiplier. In March 1890 he read to the Australian Economic Association a paper on The Law of the Incomes (1890), in which he noted that incomes arising from primary production led to an increase in income in other sectors. Using production data, and taking into account leakages abroad, he concluded that, as a general rule, incomes of primary producers equalled incomes of secondary producers; the original primary income, in other words, had a general tendency to multiply by a factor of two. De Lissa later argued that the relationship between primary and secondary income would diminish progressively until the additional income reached zero.

An area where Australia was clearly at the forefront of work internationally by the end of the 19th century was the official collection and interpretation of economic and social statistics. The most acclaimed of the colonial statisticians was Timothy Coghlan, the New South Wales Statistician, who pioneered the measurement of the national income using income, output and expenditure methods, an approach similar in many ways to modern national income accounting. Coghlan later worked in London as Agent-General for New South Wales. There he wrote a four-volume economic history of Australia – *Labour and Industry in Australia* (1918) – that drew upon quantitative information he had assembled when he was in Sydney. Later work in Australia by Colin Clark (1940), H.W. Arndt (1949), N.G. Butlin (1962) and G.D. Snooks (1994) acknowledged the ground-breaking statistical work, including national income estimation, of Coghlan and other 19th-century colonial statisticians.

Economics in the Universities

When the first universities were established in Sydney in 1851 and in Melbourne in 1854, economics was not a subject that attracted much attention. At the University of Sydney, the Professor of Classics (John Woolley) and the Professor of Philosophy (Francis Anderson) took occasional classes in economics. The Professor of Mathematics (Morris Birbeck Pell) and a later Professor of Classics (Walter Scott) gave some lectures in economics outside the university. But, as a result of growing interest in the subject by business organizations, chambers of commerce, and professional associations of bankers and accountants, courses in economics over three years began at the University of Sydney in the early 1900s. A department of economics was established in 1912, to which R.F. Irvine was appointed Professor of Economics, the first separate chair of economics in Australasia. A graduate of Canterbury University College, New Zealand, Irvine had been a pupil of James Hight. Earlier, at the University of Melbourne, Hearn had taught courses in economics for both the BA and the MA. His successor, J.S. Elkington, however, seems not to have taken the same interest in economics, and as a consequence the subject languished for a time in Melbourne.

A final year course in political economy for the BA had been offered at the University of Tasmania since the university's creation in 1889. Later a lectureship in philosophy and economics was established, but the lecturer taught courses mainly in philosophy rather than in economics. The major breakthrough in Tasmania – and, as it turned out, for economics in Australia – occurred in 1917 when Douglas Copland was appointed lecturer in history and economics. In 1920 he was appointed to a chair in economics, and later was elevated to the deanship of a new Faculty of Economics and Commerce. Like Irvine, Copland was a graduate of Canterbury University College, where he, too, had been a pupil of Hight's. In 1924 Copland was the leading force behind the establishment of the Economic Society of Australia and New Zealand, which, in the following year, published the first issue of its journal, *The Economic Record*. In the same year, 1925, Copland was appointed Professor of Commerce in the University of Melbourne.

In the University of Adelaide, founded in 1874, courses in political economy were taught by William Mitchell in the 1890s, and by Herbert Heaton in the early 1920s; in 1929 L.G. Melville was appointed to the foundation chair of economics. Meanwhile, the universities of Queensland and Western Australia, founded just before the First World War, had established combined chairs of history and economics; Henry Alcock was appointed to the chair at Queensland, and Edward Shann to the chair at the University of Western Australia. In New Zealand by the early 1920s, chairs in economics had been established at four universities: Auckland (Horace Belshaw), Canterbury (J.B. Condliffe), Otago (A.G.B. Fisher) and Wellington (Barney Murphy).

In 1914 Irvine wrote: 'When one considers the political and economic evolution of Australia, one cannot but be astonished at the neglect of these studies [that is, economics] in Australian universities' (Goodwin 1966: 636). That was certainly true of Australia prior to the First World War, but it was not true of New Zealand. By the 1890s, economics had become an important subject of study at Canterbury. There, James Hight was the foundation Professor of History and Economics. More a political historian than an economist, Hight nevertheless promoted economics as a significant field of study. A number of able students were attracted to the subject, including the first two professors of economics in Australia. By the 1920s, John Maynard Keynes could justly write that training in economics at Canterbury 'was as good as any place in the world' (Harper 1986, p. 41).

The Golden Age of Australian Economics

Yet it was in Hobart where the so-called golden age of Australian economics had its origins. Soon after his arrival at the University of Tasmania, Copland became a protégé of L.F. Giblin, a graduate in mathematics of King's College, Cambridge. Born in Tasmania, Giblin had fought on the western front in the First World War, and on leave in England had met Keynes through mutual

friends. When he returned to Hobart, Giblin was appointed Tasmanian Statistician. As a member of the Council of the University of Tasmania, he was instrumental in Copland's appointment to the newly established chair in economics and for the creation of the Faculty of Economics and Commerce. Copland then attracted J.B. Brigden to fill the lectureship that he had vacated. Copland's star pupil at Hobart was Roland Wilson, who later completed doctorates in economics at Oxford and Chicago. Wilson was to become Commonwealth Statistician and later head of the Australian Treasury. The four – Giblin, Copland, Brigden and Wilson – were at the centre of the most important work undertaken in economics in Australia from the 1920s to the 1940s.

The early promise of this group, and the coming of age of Australian economics, can be seen in Copland's paper, 'Currency Inflation and Price Movements in Australia', published in the *Economic Journal* in 1920. Using Australian data for 1901–17, and invoking Fisher's equation of exchange, Copland derived P as a residual after applying data for M, V and T. He then compared an actual price series with the hypothetical series for P, showing that the two series exhibited close agreement. Copland concluded that the 'equation of exchange may be regarded as true for Australia'. Keynes praised Copland for this work, referring as he did to Copland's 'masterly article' (Coleman et al. 2006, p. 51).

Later in the 1920s, Giblin, Copland and Brigden were appointed to the committee of enquiry into the Australian tariff *(The Australian Tariff: An Economic Enquiry,* often known as the Brigden Report) established by the federal government in 1927 (Brigden et al. 1929). The *Enquiry* concluded that, in Australian circumstances, protection had raised the 'standard of living'. This controversial conclusion, and the analysis upon which it was based, is said to have been significant for the emergence of modern international trade theory (Coleman et al. 2006, 65–73); Keynes adjudged that the *Enquiry* was 'a brilliant effort of the highest interest' (Millmow 2005, p. 1013). Similarly, Giblin's inaugural lecture in April 1930, upon his appointment to the first research chair in economics in Australia (the Ritchie Chair in the University of Melbourne), in which he produced a multiplier based on the repercussions of a decline in exports on total domestic output, is thought to have been an important stepping-stone to the eventual formulation of the Cambridge multiplier. When Giblin sent an early version of his multiplier to Keynes in August 1929, Keynes admitted that Giblin's 'method of argument' was 'novel' (Coleman et al. 2006, p. 83).

The youngest member of 'Giblin's Platoon', Roland Wilson, published a book in 1931 that attracted the attention of Viner, Harrod, Hicks, Robertson and Pigou. In *Capital Imports and the Terms of Trade,* Wilson disputed Mill's contention that the import of capital would improve a borrowing country's terms of trade. More importantly, Wilson focused on the consequences of capital imports for the price ratio of tradables to non-tradables. He showed that the ratio would decline. This conclusion was taken up in the 1970s, when it was incorporated in notions such as the Dutch disease and the Gregory thesis (named after R.G. Gregory, an Australian economist who argued in the 1970s that Australia's massive export of minerals would serve to push up the Australian dollar exchange rate with adverse consequences for other industries, particularly manufacturing industry in Australia).

Giblin's group, supported by other economists, played a decisive role in furnishing advice to Australian governments and banks during the early 1930s. The economists were critical of the central bank's policy to retain a fixed rate of exchange with sterling, advising the Bank of New South Wales early in 1931 that it should use its power and prestige as Australia's largest and oldest commercial bank to devalue the Australian pound. The economists' advice was accepted and the Australian pound was devalued. The federal and state governments then appointed Copland and Giblin to a committee (the 'Copland Committee') charged with the responsibility of formulating policies to deal with the depression. The committee's recommendations formed the core of measures included in the famous Premiers' Plan of 1931. A common theme running through the anti-depression measures proposed by Australian economists was that the loss of income

occasioned by the decline in exports should be spread among all income groups and not be confined to export and related trades. Their work was highly praised by foreign observers. Keynes, for example, wrote in 1932 that: 'I am sure that the Premiers' Plan last year saved the economic structure of Australia' (1932, p. 94). As a measure of the influence of Australian economists, Copland was invited to present the inaugural Alfred Marshall Memorial Lectures in Cambridge in 1933; the lectures were published under the title *Australia in the World Crisis, 1929–1933* (1934).

Australian economists were prominent again during and immediately after the Second World War. Shortly before the outbreak of war, the federal government established an Economic and Financial Committee (the F&E) to advise it on economic questions that might arise in the event of war. Giblin was appointed chairman of the committee, which included Copland, Brigden and Wilson. When the war came, the F&E formulated the government's approach to war finance, following principles that Keynes had put to the British government.

When it came to formulating plans for post-war reconstruction, Australian economists prepared at the government's request a domestic employment policy based on demand management. Their proposals were published in the famous government white paper of 1945, *Full Employment in Australia* (Cornish 1981). The economists supported Keynes's Clearing Union, opposing as they did the rival Stabilization Fund of the United States Treasury. In fact, they went further than Keynes by formulating what they called the 'international full employment approach' or 'positive approach', sometimes known as 'Australia's Keynesian crusade' (Cornish 1993). This policy arose from Article VII of the Mutual Aid Agreement signed in 1942. In return for United States assistance during the war, recipient countries pledged to enter discussions aimed at liberalizing foreign trade and international payments. Given uncertainty about the restoration of world trade, and concerned about the impact on employment of abolishing preferential trade arrangements, the 'positive approach' maintained that Australia would support Article VII provided the United States and other major economic powers committed themselves to policies aimed at maintaining full employment in their domestic economies. Such policies, it was believed, would provide a buoyant demand for Australian exports. Australian representatives promoted the 'positive approach' at major international conferences during the 1940s, including those at Bretton Woods, San Francisco and Havana.

Australasian Economics Since the Second World War

The numbers working in economics increased enormously after the Second World War. It is estimated that, in Australia, whereas 5000 persons graduated in economics between 1916 and 1947, 50,000 graduated between 1947 and 1986 (Butlin 1987). While there had been no increase in Australian universities between the two world wars, between 1945 and the early 1990s the number rose from six to more than 30. Some of the newer universities offered economics simply as a subsidiary course in business studies programmes; most, however, offered specialist degrees in economics (Groenewegen 1996). In the 1970s, reflecting the growth of economists, the Economics Society of Australia and New Zealand was divided into two professional organizations – the Economic Society of Australia, and the New Zealand Economic Association. Yet another indicator of the expanding scale of the discipline was the increase in the number of journals dedicated to economics, from one in 1945 (*Economic Record*) to four by the mid-1960s (the additions were *Australian Economic Papers, Australian Economic Review* and *New Zealand Economic Papers*).

However distinctive the character of Australasian economics may have been in the interwar period, it disappeared after the Second World War as the American approach, with its emphasis on model building, mathematics and econometrics, began to dominate the discipline (Groenewegen and McFarlane 1990). It is understandable perhaps that economists seeking to publish their work in leading international journals, many of them

American-based, would want to incorporate the latest ideas and methods arising in the United States. The Americanization of the discipline also stemmed in part from the increasing number of students from Australasia going to the United States for postgraduate studies; previously the United Kingdom (Cambridge in particular) had been the destination for graduate studies in economics. Yet the American dominance of economics did not inhibit Australian and New Zealand economists from making important contributions to the subject. For example, there was the work of T.W. Swan (1956, 1963) and W.E.G. Salter (1959) in growth theory and on issues of internal–external balance in small dependent economies; W.M. Corden's work in the theory and measurement of effective protection, tariff policy and international monetary economics (1971); Murray Kemp's formulation of general equilibrium trade models (1964); G.C. Harcourt's writing on capital theory (1986); A.W. Phillips's contributions to the theory and measurement of inflation, and the relation between wages and unemployment (1958); and the writing on Australia–Asia economic relations by J.G. Crawford (Evans and Miller 1987), H.W. Arndt (1972) and Ross Garnaut (2001).

See Also

- ▶ Arndt, Heinz Wolfgang (Born 1915)
- ▶ Butlin, Noel George (1921–1991)
- ▶ Clark, Colin Grant (1905–1989)
- ▶ Swan, Trevor W. (1914–1989)

Bibliography

Arndt, H.W. 1949. A pioneer in national income estimates. *Economic Journal* 59: 616–625.

Arndt, H.W. 1972. *Australia and Asia: Economic essays*. Canberra: Australian National University Press.

Brigden, J.B., D.B. Copland, E.C. Dyason, L.F. Giblin, and C.H. Wickens. 1929. *The Australian tariff: An economic enquiry*. Melbourne: Melbourne University Press.

Butlin, N.G. 1962. *Australian domestic product, investment and foreign borrowing 1861–1938/39*. Cambridge: Cambridge University Press.

Butlin, N.G. 1987. *Human or inhuman capital? The economics profession 1916–87*. Working Papers in Economic History No. 91, Australian National University.

Clark, C. 1940. *Conditions of economic progress*. London: Macmillan.

Coghlan, T.A. 1918. *Labour and industry in Australia from the first settlement in 1788 to the establishment of the Commonwealth in 1901*. London and New York: Oxford University Press.

Coleman, W., S. Cornish, and A. Hagger. 2006. *Giblin's platoon: The trials and triumph of the economist in Australian public life*. Canberra: Australian National University Press.

Copland, D.B. 1920. Currency inflation and price movements in Australia. *Economic Journal* 30: 484–505.

Copland, D.B. 1934. *Australia in the world crisis, 1929–1933*. Cambridge: Cambridge University Press.

Corden, W.M. 1971. *The theory of protection*. Oxford: Clarendon Press.

Cornish, S. 1981. *Full employment in Australia: The genesis of a white paper*. Research Papers in Economic History No. 1. Canberra: Australian National University.

Cornish, S. 1993. The Keynesian revolution in Australia: Fact or fiction? *Australian Economic History Review* 33(2): 42–68.

De Lissa, A. 1890. The law of the incomes. *Australian Economist* 2(1): 6–17.

Endres, A.M. 1991. J.B. Condliffe and the early Canterbury tradition in economics. *New Zealand Economic Papers* 25(2): 171–197.

Evans, L.T., and J.D.B. Miller, ed. 1987. *Policy and practice: Essays in honour of Sir John Crawford*. Canberra: Australian National University Press.

Garnaut, R. 2001. *Social democracy in Australia's Asian future*. Canberra: Asia Pacific Press.

Giblin, L.F. 1930. *Australia 1930*. Melbourne: Melbourne University Press.

Goodwin, C.D.W. 1966. *Economic enquiry in Australia*. Durham, NC: Duke University Press.

Groenewegen, P. 1996. The Australian experience. In *The Post-1945 Internationalization of Economics*, ed. A.-W. Coats. Durham, NC and London: Duke University Press. Annual Supplement to *History of Political Economy* 28.

Groenewegen, P., and B. McFarlane. 1990. *A history of Australian economic thought*. London and New York: Routledge.

Harcourt, G.C. 1972. *Some Cambridge controversies in the theory of capital*. Cambridge: Cambridge University Press.

Harper, M. 1986. Melbourne economists in the public arena: From Copland to the Institute. In *Victoria's heritage. Lectures to celebrate the 150th anniversary of European settlement in Victoria*, ed. A.G.L. Shaw. Sydney: Allen & Unwin.

Hearn, W.E. 1863. *Plutology*. Melbourne: George Robertson.

Kemp, M. 1964. *The pure theory of international trade*. Englewood Cliffs: Prentice Hall.
Keynes, J.M. 1932. Report of the Australian experts. In *The collected writings of John Maynard Keynes*, ed. D. Moggridge, Vol. 21, 94–100. London: Macmillan and Cambridge University Press for the Royal Economic Society 1982.
Lang, J.D. 1834. *An historical and statistical account of New South Wales*, 1852. London: Longman, Brown, Green and Longmans.
Millmow, A. 2005. Australian economics in the twentieth century. *Cambridge Journal of Economics* 29: 1011–1026.
Musgrave, A. 1875. *Studies in political economy*. London: Henry S. King and Co..
Nadel, G. 1957. *Australia's colonial culture. Ideas, men and institutions in mid-nineteenth century eastern Australia*. Melbourne: F.W. Cheshire.
Phillips, A.W. 1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* 25: 283–299.
Salter, W.E.G. 1959. Internal and external balance: The role of price and external expenditure effects. *Economic Record* 35: 226–238.
Snooks, G.D. 1994. *Portrait of the family within the total economy: A study in longrun dynamics, Australia 1788–1990*. Cambridge: Cambridge University Press.
Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32: 334–361.
Swan, T.W. 1963. Longer run problems of the balance of payments. In *The Australian economy: A volume of readings*, ed. H.W. Arndt and W.M. Corden. Melbourne: F.W. Cheshire.
Syme, D. 1876. *Outlines of an industrial science*. London: Henry S. King and Co..
Wentworth, W.C. 1819. *A statistical, historical and political description of the colony of New South Wales and its dependent settlements in Van Diemen's Land*. London: G. & W.B. Whittaker.
Wilson, R. 1931. *Capital imports and the terms of trade, examined in the light of sixty years of Australian borrowings*. Melbourne: Melbourne University Press.

Austrian Economics

Israel M. Kirzner

Keywords

Austrian capital theory; Austrian economics; Austrian School; Austrian theory of the business cycle; Böhm-Bawerk, E. von; Exploitation; German Historical School; Haberler, G.; Hayek, F. A. von; Higher-order goods; Jaffé, W.; Jevons, W. S.; Labour theory of value; Lange, O. R.; Libertarianism; Machlup, F.; Marginal productivity theory; Marginal revolution; Marginal utility; Marginalism; Market process; Marxist economics; Mayer, H.; Menger, C.; Methodenstreit; Methodological individualism; Methodological subjectivism; Mises, L. E. von; Morgenstern, O.; Non-market prices; Opportunity cost; Robbins, L. C.; Rosenstein-Rodan, P. N.; Rothbard, M. N.; Roundabout production processes; Schams, E.; Schonfield, L.; Schumpeter, J. A.; Shackle, G. L. S.; Socialist calculation debate; Strigl, R.; Subjective theory of value; Surplus value; Time preference; Uncertainty; Utility; Value and price; Walras, L.; Wicksell, J. G. K.; Wieser, F. F. von

JEL Classifications

B2

The birth of the Austrian School of economics is usually recognized as having occurred with the 1871 publication of Carl Menger's *Grundsätze der Volkwirthschaftslehre.* On the basis of this work Menger (hitherto a civil servant) became a junior faculty member at the University of Vienna. Several years later, after a stint as tutor and travelling companion to Crown Prince Rudolph, he was appointed to a professorial chair at the University. Two younger economists, Eugen von Böhm-Bawerk and Friedrich von Wieser (neither of whom had been a student of Menger), became enthusiastic supporters of the new ideas put forward in Menger's book. During the 1880s a vigorous outpouring of literature from these two followers, from several of Menger's students, and in particular a methodological work by Menger himself, brought the ideas of Menger and his followers to the attention of the international community of economists. The Austrian School was now a recognized entity. Several works of Böhm-Bawerk and Wieser were translated into English; and by 1890 the editors of the

US journal *Annals of the American Academy of Political and Social Science* were asking Böhm-Bawerk for an expository paper explaining the doctrines of the new school. What follows seeks to provide a concise survey of the history of the Austrian School with special emphasis on (*a*) the major representatives of the school; (*b*) the central ideas identified with the school; (*c*) the relationship between the school and its ideas, and other major schools of thought within economics; (*d*) the various meanings and perceptions associated today with the term Austrian economics.

The Founding Austrians

Menger's (1871) book is recognized in the history of economic thought (alongside Jevons's 1871 *Theory of Political Economy,* and Walras's 1874 *Eléments d'économie politique pure*) as a central component of the 'marginalist revolution'. For the most part, historians of thought have emphasized the features in Menger's work that parallel those of Jevons and Walras. More recently, following especially the work of W. Jaffé (1976) attention has come to be paid to those aspects of Menger's ideas which set them apart from those of his contemporaries. A series of recent studies (Grassl and Smith 1986) have related these unique aspects of Menger and the early Austrian economists to broader currents in the late 19th-century intellectual and philosophical scene in Austria.

The central thrust of Menger's book was unmistakable; it was an attempt to rebuild the foundations of economic science in a way which, while retaining the abstract, theoretical character of economics, offered an understanding of value and price which ran sharply counter to classical teachings. For the classical economists value was seen as governed by past resource costs; Menger saw value as expressing judgements concerning future usefulness in meeting consumer wants. Menger's book, offered to the German-speaking scholarly community of Germany and Austria, was thus altogether different, in approach, style and substance, from the work coming from the German universities. That latter work, while also sharply critical of classical economics, was attacking its theoretical character, and appealing for a predominantly historical approach. At the time Menger's book appeared, the 'older' German Historical School (led by Roscher, Knies and Hildebrand) was beginning to be succeeded by the 'younger' Historical School, whose leader was to be Gustav Schmoller. Menger, the 31-year-old Austrian civil servant, was careful not to present his work as antagonistic to that of German economic scholarship. In fact he dedicated his book – with 'respectful esteem' – to Roscher, and offered it to the community of German scholars 'as a friendly greeting from a collaborator in Austria and as a faint echo of the scientific suggestions so abundantly lavished on us Austrians by Germany . . .' (Menger 1871, Preface). Clearly Menger hoped that his theoretical innovations might be seen as reinforcing the conclusions derived from historical studies of the German scholars, contributing to a new economics to replace a discredited British classical orthodoxy.

Menger was to be bitterly disappointed. The German economists virtually ignored his book; where it was noticed in the German language journals it was grossly misunderstood or otherwise summarily dismissed. For the first decade after the publication of his book, Menger was virtually alone; there was certainly no Austrian 'school'. And when the enthusiastic work of Böhm-Bewark and Wieser began to appear in the 1880s, the new literature acquired the appellation 'Austrian' more as a pejorative epithet bestowed by disdainful German economists than as an honorific label (Mises 1969, p. 40). This rift between the Austrian and German scholarly camps deepened most considerably after the appearance of Menger's methodological challenge to the historical approach (Menger 1883). Menger apparently wrote that work having been convinced by the unfriendly disinterest with which his 1871 book had been received in Germany, that German economics could be rescued only by a frontal attack on the Historical School. The bitter *Methodenstreit* that followed is usually (but not invariably, see Bostaph 1978) seen by historians of economics as constituting a tragic waste of scholarly energy. Certainly this

venomous academic conflict helped bring the existence of an Austrian School to the attention of the international economics fraternity – as a group of dedicated economists offering a flood of exciting theoretical ideas reinforcing the new marginalist literature, sharply modifying the hitherto dominant classical theory of value. Works by Böhm-Bawerk (1886), Wieser (1884, 1889), Komorzynski (1889) and Zuckerkandl (1889) offered elaborations or discussions of Menger's central, subjectivist ideas on value, cost, and price. Works on the theory of pure profit, and on such applications as public finance theory, were contributed by writers such as Mataja (1884), Gross (1884), Sax (1887), and Meyer (1887). The widely used textbook by Philippovich (1893), who was a professor at the University of Vienna (but more sympathetic towards the contributions of the German School), is credited with an important role in spreading Austrian marginal utility theory among German-language students.

In these early Austrian contributions to the theory of value and price, emphasis was (as in the Jevonsian and Walrasian approaches) placed both on marginalism and on utility. But important differences set the Austrian theory apart from other early marginalist theories. The Austrians made no attempt to present their ideas in mathematical form, and as a consequence the Austrian concept of the margin differs somewhat from that of Jevons and Walras. For the latter, and for subsequent microeconomic theorists, the marginal value of a variable refers to the instantaneous rate of change of the 'total' variable. But the Austrians worked, deliberately, with discrete variables (see Menger 1973). More importantly the concept of marginal utility, and the sense in which it decreases, referred for the Austrians not to psychological enjoyments themselves, but to (ordinal) marginal *valuations* of such enjoyments (McCulloch 1977). In any event, as has been urged by Streissler (1972), what was important for the Austrians in marginal utility was not so much the adjective as the noun. Menger saw his theory as demonstrating the unique and exclusive role played, in the determination of economic value, by subjective, 'utility', considerations. Values are not seen (as they are in Marshallian economics) as *jointly* determined by subjective (utility) and objective (physical cost) considerations. Rather values are seen as determined *solely* by the actions of consumers (operating within a given framework of existing commodity and/or production possibilities). Cost is seen (by Menger, and especially by Wieser, whose name came to be associated closely with this insight) merely as prospective utility deliberately sacrificed (in order to command more highly preferred utility). Whereas in the development of the other marginalist theories, it took perhaps two decades for it to be seen that marginal utility value theory points directly to marginal productivity distribution theory. Menger at least glimpsed this insight immediately. His theory of 'higher-order' goods emphasizes how both the economic character and the value of factor services are derived exclusively from the valuations placed by consumers upon the consumers products to whose emergence these higher-order goods ultimately contribute. Böhm-Bawerk contributed not only to the exposition and dissemination of Menger's basic subjective value theory, but most prominently also to the theory of capital and interest. Early in his career he published a massive volume (Böhm-Bawerk 1884) in the history of doctrine, offering an encyclopedic critique of all earlier theories of interest (or 'surplus value' or 'normal profit'). This he followed up several years later with a volume (Böhm-Bawerk 1889) presenting his own theory. At least part of the renown of the Austrian School at the turn of the century derived from the fame of these contributions. As we shall note later on, a number of subsequent and modern writers (such as Hicks 1973; Faber 1979; Hausman 1981) have indeed seen these Böhm-Bawerkian ideas as constituting the enduring element of the Austrian contribution. Others, taking their cue from an oft-repeated critical remark attributed to Menger (Schumpeter 1954, p. 847 n. 8), have seen Böhm-Bawerk's theory of capital and interest as separate from, or even as somehow inconsistent with, the core of the Austrian tradition stemming from Menger (Lachmann 1977, p. 27). Certainly Böhm-Bawerk himself saw his theory of capital and interest as a seamless extension of basic subjectivist value theory. Once the dimension of time

has been introduced into the analysis of both consumer and producer decisions, Böhm-Bawerk found it possible to explain the phenomenon of interest. Because production takes time, and because economizing men systematically choose earlier receipts over (physically similar) later receipts, capitalusing production processes cannot fail to yield (even after the erosive forces of competition are taken into account) a portion of current output to those who in earlier periods invested inputs into time-consuming, 'roundabout' production processes.

Böhm-Bawerk became, indeed, so prominent a representative of the Austrian School prior to World War I that, largely due to his work, the Marxists came to view the Austrians as the quintessential bourgeois, intellectual enemy of Marxist economics (Bukharin 1914). Not only did Böhm-Bawerk offer his own theory explaining the phenomenon of the interest 'surplus' in a manner depriving this capitalist income of any exploitative character, he had emphatically and mercilessly refuted Marxist theories of this surplus. In his 1884 work Böhm-Bawerk had systematically deployed the Austrian subjective theory of value to criticize witheringly the Marxist labour theory underlying the exploitation theory. A decade later (Böhm-Bawerk 1896) he offered a patient, but relentless and uncompromising elaboration of that critique (in dissecting the claim that Marx's posthumously published volume 3 of *Capital* could be reconciled with the simple labour theory forming the basis of volume 1). This tension between the Marxists and the Austrians was to find later echoes in the debate which Mises and Hayek (third- and fourth- generation Austrians) were to conduct, during the 1920–40 interwar period, with socialist economists concerning the possibility of economic calculation in a centrally planned economy.

Menger retired from his University of Vienna professorship in 1903. His chair was assumed by Wieser. Wieser has been justly described as

> the central figure of the Austrian School: central in time, central in the ideas he propounded, central in his intellectual abilities, that is to say neither the most outstanding genius nor one of those also to be mentioned . . . He had the longest teaching record . . . (Streissler 1986)

Wieser had been an early and prolific expositor of Menger's theory of value. His general treatise on economics, summing up his life's contributions (Wieser 1914), has been hailed by some (but certainly not all) commentators as a major achievement. (Hayek 1968, sees the work as a personal achievement rather than as representative of the Austrian School.) In the decade prior to the First World War, it was Böhm-Bawerk's seminar (begun when Böhm-Bawerk rejoined academic life after a number of years as Finance Minister of Austria) that became famous as the intellectual centre of the Austrian School. Among the subsequently famous economists who participated in the seminar were Josef A. Schumpeter and Ludwig von Mises, both of whom published books prior to the war (Schumpeter 1908, 1912; Mises 1912).

After the First World War

The scene in Austrian economics after the war was rather different than it had been before. Böhm-Bawerk had died in 1914. Menger, who even in his long seclusion after retirement, used to receive visits from the young economists at the university, died in 1921. Although Wieser continued to teach until his death in 1926, the focus shifted to younger scholars. These included particularly Mises, the student of Böhm-Bawerk, and Hans Mayer, who succeeded his teacher Wieser, to his chair. Mises, although an 'extraordinary' (unsalaried) faculty member at the university, never did obtain a professorial chair. Much of his intellectual influence was exercised outside the university framework (Mises 1978, ch. ix). Other notable (pre-war-trained) scholars during the 1920s included Richard Strigl, Ewald Schams, and Leo Schonfeld (later Illy). In the face of these changes the Austrian tradition thrived. New books were published, and a new crop of younger students came to the fore, many of whom were to become internationally famous economists in later decades. These included particularly Friedrich A. Hayek, Gottfried Haberler, Fritz Machlup, Oskar Morgenstern, and Paul N. Rosenstein-Rodan. Economic discussion

among the Austrians was vigorously carried on, during the 1920s and early 1930s, within two partly overlapping groups. One, at the university, was led by Hans Mayer. The other centred on Mises, whose famed *privatseminar* met in his Chamber of Commerce office and drew not only the gifted younger economists, but also such philosophers, sociologists and political scientists as Felix Kaufmann, Alfred Schutz and Erik Voegelin. It was during this period that British economist Lionel Robbins came decisively under the influence of the intellectual ferment going on in Vienna. A distinctly important outcome of this contact was Robbins's highly influential book (Robbins 1932). It was largely through this work that a number of key Austrian ideas came to be absorbed into the mainstream literature of 20th-century Anglo-American economics. In 1931 Robbins invited Hayek to lecture at the London School of Economics, and this led to Hayek's appointment to the Tooke chair at that institution.

Hayek's arrival on the British scene contributed especially to the development and widespread awareness of the 'Austrian' theory of the business cycle. Mises had sketched such a theory as early as 1912 (Mises 1912, pp. 396–404). This theory attributed the boom phase of the cycle to intertemporal misallocation stimulated by 'too low' interest rates. This intertemporal misallocation consisted of producers initiating processes of production that implicitly anticipated a willingness on the part of the public to postpone consumption to a degree in fact inconsistent with the true pattern of time preferences. The subsequent abandonment of unsustainable projects constitutes the down phase of the cycle. Mises emphasized the roots of this theory in Wicksell, and in earlier insights of the British Currency School. Indeed Mises was tempted to challenge the appropriateness of the 'Austrian' label widely attached to the theory (Mises 1943). But, as he recognized, the Austrian label had become firmly attached to the doctrine. Hayek's vigorous exposition and extensive development of the theory (Hayek 1931, 1933, 1939) and his introduction (through the theory) of Böhm-Bawerkian capital-theoretic insights to the British public, unmistakably left Hayek's imprint on the fully developed theory, and taught the profession to see it as a central contribution of the Austrian School. Given all these developments it is apparent that we must consider the early 1930s as constituting in many ways the period of greatest Austrian School influence upon the economics profession generally. Yet this triumph was to be short-lived indeed.

With the benefit of hindsight it is perhaps possible to understand why and how this same period of the early 1930s constituted, in fact, a decisive, almost fatal, turning point in the fortunes of the School. Within a few short years the idea of a distinct Austrian School – except as an important, but bygone, episode in the history of economics – virtually disappeared from the economics profession. While Hans Mayer continued to occupy his chair in Vienna until after the Second World War, the group of prominent younger economists who had surrounded Mises soon dispersed (for political or other reasons), many of them to various universities in the United States. With Mises migrating in 1934 to Geneva and later to New York, with Hayek in London, Vienna ceased to be a centre for the vigorous continuation of the Austrian tradition. Moreover, many of the group were convinced that the important ideas of the Austrian School had now been successfully absorbed into mainstream economics. The emerging ascendancy of theoretical economics, and thus the eclipse of historicist and anti-theoretical approaches to economics, no doubt permitted the Austrians to believe that they had finally prevailed, that there was no longer any particular need to cultivate a separate Austrian version of economic theory. A 1932 statement by Mises captures this spirit. Referring to the usual separation of economic theorists into three schools of thought, 'the Austrian and the Anglo-American Schools and the School of Lausanne', Mises (citing Morgenstern) emphasized that these groups 'differ only in their mode of expressing the same fundamental idea and that they are divided more by their terminology and by peculiarities of presentation than by the substance of their teachings' (Mises 1933, p. 214). Yet the survival and development of an Austrian tradition

during and subsequent to the Second World War, largely through the work of Mises himself and of Hayek, deserves and requires attention.

Fritz Machlup has, on several occasions (Machlup 1981, 1982) listed six ideas as central to the Austrian School prior to the Second World War. There is every reason to agree that it was these six ideas that expressed the Austrian approach as understood, say, in 1932. These ideas were: (*a*) methodological individualism (not to be confused with political or ideological individualism, but referring to the claim that economic phenomena are to be explained by going back to the actions of individuals); (*b*) methodological subjectivism (recognizing that the actions of individuals are to be understood only by reference to the knowledge, beliefs, perception and expectations of these individuals); (*c*) marginalism (emphasizing the significance of prospective *changes* in relevant magnitudes confronting the decision maker); (*d*) the influence of utility (and diminishing marginal utility) on demand and thus on market prices; (*e*) opportunity costs (recognizing that the costs that affect decisions are those that express the most important of the alternative opportunities being sacrificed in employing productive services for one purpose rather than for the sacrificed alternatives); (*f*) time structure of consumption and production (expressing time preferences and the productivity of 'roundaboutness').

It seems appropriate, however, to comment further on this list. (1) With varying degrees of emphasis most modern microeconomics incorporates all of these ideas, so that (2) this list supports the cited Morgenstern–Mises statement emphasizing the common ground shared by *all* schools of economic theory. However (3) subsequent developments in the work of Mises and Hayek suggest that the list of six Austrian ideas was not *really* complete. While few Austrians at the time (of the early 1930s) were perhaps able to identify additional Austrian ideas, such additional insights were in fact implicit in the Austrian tradition and were to be articulated explicitly in later work. From this perspective, then, (4) important *differences* separate Austrian economic theory from the mainstream developments in microeconomics, particularly as these latter developments proceeded from the 1930s onwards. It was left for Mises and Hayek to articulate these differences and thus preserve a unique Austrian 'presence' in the profession.

Later Developments in Austrian Economics

One early expression of such differences between the Austrian understanding of economic theory and that of other schools, was Hans Mayer's paper criticizing 'functional price theories' and calling for the 'genetic-causal' method (Mayer 1932). Here Mayer was criticizing equilibrium theories of price that neglected to explicate the *sequence* of actions leading to market prices. To understand this sequence one must understand the causal genesis of the component actions in the sequence. In the light of the later writings of Mises and Hayek, it seems reasonable to recognize Mayer as having placed his finger on an important and distinctive element embedded in the Austrian understanding. Yet the Austrians themselves during the 1920s (and such students of their works as Lionel Robbins) seemed to have missed this insight. What appears to have helped Hayek and Mises articulate this hitherto overlooked element was the well-known interwar debate concerning the possibility of economic calculation under central planning. A careful reading of the contributions to that debate suggests that it was in reaction to the 'mainstream' equilibrium arguments of their opponents that Mises and Hayek made explicit the emphasis on process, learning and discovery to be found in the Austrian understanding of markets (Lavoie 1985).

Mises had argued that economic calculation calls for the guidance supplied by prices; since the centrally planned economy has no market for productive factors, it cannot use factor prices as guides. Oskar Lange and others countered that prices need not be market prices; that guidance could be provided by non-market prices, announced by the central authorities, and treated by socialist managers 'parametrically' (just as prices are treated by producers in the theory of

the firm, in perfectly competitive factor and product markets). It was in response to this argument that Hayek developed his interpretation of competitive market processes as processes of discovery during which dispersed information comes to be mobilized (Hayek 1949, chs. 2, 4, 5, 7, 8, 9). An essentially similar characterization of the market process (without the Hayekian emphasis on the role of knowledge, but with an accent on entrepreneurial activity in a world of open-ended, radical uncertainty) was presented by Mises during the same period (Mises 1940, 1949). In the light of these Mises–Hayek developments in the theory of market process (and recognizing that these developments constituted the articulation of insights taken for granted in the early Austrian tradition: Kirzner 1985; Jaffé 1976), it seems reasonable to add the following to Machlup's list of ideas central to the Austrian tradition: (*g*) markets (and competition) as processes of learning and discovery; (*h*) the individual decision as an act of choice in an essentially uncertain context (where the identification of the relevant alternatives is part of the decision itself). It is these latter ideas that have come to be developed in and made central to the revived attention to the Austrian tradition that, stemming from the work of Mises and Hayek, has emerged in the United States in recent decades.

Austrian Economics Today

As a result of these somewhat varied developments in the history of the Austrian School since 1930, the term 'Austrian economics' has come to evoke a number of different connotations in contemporary professional discussion. Some of these connotations are, at least partly, overlapping; others are, at least partly, mutually inconsistent. If seems useful, in disentangling these various perceptions, to identify a number of different meanings that have come to be attached to the term 'Austrian economics' in the 1980s. The present status of the Austrian School of economics is, for better or for worse, encapsulated in these current perceptions.

1. For many economists the term 'Austrian economics' is strictly a historical term. In this perception the existence of the Austrian School did not extend beyond the early 1930s: Austrian economics was partly absorbed into mainstream microeconomics, and partly displaced by emerging Keynesian macroeconomics. To a considerable extent this view seems to be that held by economists in Austria today. Economists (and other intellectuals) in Austria today are thoroughly cognizant of – and proud of – the earlier Austrian School, as evidenced by several commemorative conferences held in Austria in recent years, and by several related volumes (Hicks and Weber 1973; Leser 1986), but see themselves today simply as a part of the general community of professional economists. Erich Streissler, holder of the chair occupied by Menger, Wieser and Mayer, has written extensively, and with the insights and scholarship of one profoundly influenced by the Austrian tradition, concerning numerous aspects of the Austrian School and its principal representatives (Streissler 1969, 1972, 1973, 1986).
2. For a number of economists the adjective 'Austrian' has come to mark a revival of interest in Böhm-Bawerkian capital-and-interest theory. This revival has emphasized particularly the time dimension in production and the productivity of roundaboutness. Among the contributors to this literature should be mentioned Hicks (1973), Bernholz (1971, 1973), Faber (1979) and Orosel (1981). In this literature, then, the term 'Austrian' has very little to do with the general subjectivist Mengerian tradition (which had, as noted earlier, certain reservations in regard to the Böhm-Bawerkian theory).
3. For other economists (and non-economists) the term 'Austrian economics' has come to be associated less with a unique methodology, or with specific economic doctrines, than with libertarian ideology in political and social discussion. For these observers, to be an Austrian economist in the 1980s is simply to be in favour of free markets. Machlup (1982) has noted (and partly endorsed) this perception of the term 'Austrian'. He has ascribed it,

particularly, to the impact of the work of Mises. Mises' championship of the market cause was so prominent, and his identification as an Austrian was at the same time so unmistakable, that it is perhaps natural that his strong policy pronouncements in support of unhampered markets came to be perceived as the core of Austrianism in modern times. This has been reinforced by the work of a leading US follower of Mises, Murray N. Rothbard, who was also prominent in libertarian scholarship and advocacy. Other observers, however, would question this identification. While, as earlier noted, many of the early contributions of the Austrian School were seen as sharply antagonistic to Marxian thought, the school on the whole maintained an apolitical stance. Among the founders of the school, Wieser was in fact explicit in endorsing the interventionist conclusions of the German Historical School (Wieser 1914, p. 490ff). While both Mises and Hayek provocatively challenged the possibility of efficiency under socialism, they too, emphasized the *wertfrei* character of their economics. Both writers would see their free market stance at the policy level as related to, but not as central to, their Austrianism.

4. For many in the profession the term 'Austrian economics' has come, since about 1970, to refer to a revival of interest in the ideas of Carl Menger and the earlier Austrian School, particularly as these ideas have been developed through the work of Mises and Hayek. This revival has occurred particularly in the United States, where a sizeable literature has emerged from a number of economists. This literature includes, in particular, works by Murray N. Rothbard (1962), Israel Kirzner (1973), Gerald P. O'Driscoll (1977, 1985), Mario J. Rizzo (O'Driscoll and Rizzo 1985), and Roger W. Garrison (1978, 1982, 1985). The thrust of this literature has been to emphasize the differences between the Austrian understanding of markets as processes, and that of the equilibrium theorists whose work has dominated much of modern economic theory. As a result of this emphasis, this sense of the term 'Austrian economics' has often (and only partly accurately; see White 1977, p. 9) come to be understood as a refusal to adopt modern mathematical and econometric techniques – which standard economics adopted largely as a result of its equilibrium orientation. The economists in this group of modern Austrians (sometimes called neo-Austrian) do see themselves as continuators of an earlier tradition, sharing with mainstream neoclassical economics an appreciation for the systematic outcomes of markets, but differing from it in its understanding of how these outcomes are in fact achieved. Largely as a result of the activity of this group, many classic works of the early Austrians have recently been republished in original or translated form, and have attracted a considerable readership both inside and outside the profession.
5. Yet another current meaning loosely related to the preceding sense of the term has come to be associated with the term 'Austrian economics'. This meaning refers to an emphasis on the radical uncertainty that surrounds economic decision making, to an extent that implies virtual rejection of much of received microeconomics. Ludwig Lachmann (1976) has identified the work of G.L.S. Shackle as constituting in this regard the most consistent extension of Austrian (and especially of Misesian) subjectivism. Lachmann's own work (1973, 1977, 1986) has, in the same vein, stressed the indeterminacy of both individual choices and market outcomes.

This line of thought has come to imply serious reservations concerning the possibility of systematic theoretical conclusions commanding significant degrees of generality. This connotation of the term 'Austrian economics' thus associates it with a stance sympathetic, to a degree, towards historical and institutional approaches. Given the prominent opposition of earlier Austrians to these approaches, this association has, as might be expected, been seen as ironic or even paradoxical by many observers (including, especially, modern exponents of the broader tradition of the Austrian School of economics).

[An earlier article on the Austrian School of economics was begun and substantially drafted by Professor Friedrich A. Hayek – himself a Nobel laureate in economics whose celebrated contributions are deeply rooted in the Austrian tradition. The present author gratefully acknowledges his indebtedness (in the writing of this essay) to the characteristic scholarship and treasure trove of facts contained in Professor Hayek's unfinished article, as well as to Professor Hayek's other numerous studies that relate to the history of the Austrian School.]

See Also

- ▶ Böhm-Bawerk, Eugen von (1851–1914)
- ▶ Competition, Austrian
- ▶ Hayek, Friedrich August von (1899–1992)
- ▶ Imputation
- ▶ Menger, Carl (1840–1921)
- ▶ Mises, Ludwig Edler von (1881–1973)

Bibliography

Bernholz, P. 1971. Superiority of roundabout processes and positive rate of interest. A simple model of capital and growth. *Kyklos* 24: 687–721.

Bernholz, P., and M. Faber. 1973. Technical superiority of roundabout processes and positive rate of interest. A capital model with depreciation and n-period horizon. *Zeitschrift für die gesamte Staatswissenschaften* 129: 46–61.

Bostaph, S. 1978. The methodological debate between Carl Menger and the German historicists. *Atlantic Economic Journal* 6 (3): 3–16.

Bukharin, N. 1914. *The economic theory of the Leisure class* (trans: from Russian, London: M. Lawrence, 1927; reprinted). New York: Monthly Review Press, 1972.

Faber, M. 1979. *Introduction to modern Austrian capital theory*. Berlin: Springer.

Garrison, R.W. 1978. Austrian macroeconomics: A diagrammatical exposition. In *New directions in Austrian economics*, ed. L.M. Spadaro. Kansas City: Sheed, Andrews & McMeel.

Garrison, R.W. 1982. Austrian economics as the middle ground: Comment on Loasby. In *Method, process, and Austrian economics: Essays in honor of Ludwig von Mises*, ed. I.M. Kirzner. Lexington: Lexington Books.

Garrison, R.W. 1985. Time and money: The universals of macroeconomic theorizing. *Journal of Macroeconomics* 6: 197–213.

Grassl, W., and B. Smith, eds. 1986. *Austrian economics, historical and philosophical background*. New York: New York University Press.

Gross, G. 1884. *Die Lehre von Unternehmergewinn*. Leipzig: Duncker & Humblot.

Hausman, D.M. 1981. *Capital, profits, and prices*. New York: Columbia University Press.

Hayek, F.A. 1931. *Prices and production*. London: Routledge & Sons.

Hayek, F.A. 1933. *Monetary theory and the trade cycle*. London: Jonathan Cape.

Hayek, F.A. 1939. *Profits, interest and investment: And other essays on the theory of industrial fluctuations*. London: Routledge & Kegan Paul.

Hayek, F.A. 1949. *Individualism and economic order*. London: Routledge & Kegan Paul.

Hayek, F.A. 1968. Economic thought VI: The Austrian school. In *International encyclopedia of the social sciences*, ed. D.L. Sills. New York: Macmillan.

Hicks, J. 1973. *Capital and time: A neo-Austrian theory*. Oxford: Clarendon Press.

Hicks, J.R., and W. Weber. 1973. *Carl Menger and the Austrian school of economics*. Oxford: Clarendon Press.

Jaffé, W. 1976. Menger, Jevons and Walras de-homogenized. *Economic Inquiry* 14: 511–524.

Jevons, W.S. 1871. *The theory of political economy*. London: Macmillan.

Kauder, E. 1965. *A history of marginal utility theory*. Princeton: Princeton University Press.

Kirzner, I.M. 1973. *Competition and entrepreneurship*. Chicago: University of Chicago Press.

Kirzner, I.M. 1981. Mises and the renaissance of Austrian economics. In *Homage to mises, the first hundred years*, ed. J.K. Andrews Jr. Hillsdale: Hillsdale College Press.

Kirzner, I.M. 1985. Comment on R.N. Langlois, 'From the knowledge of economics to the economics of knowledge: Fritz Machlup on methodology and on the "knowledge society"'. In *Research in the history of economic thought and methodology*, ed. Warren J. Samuels. Greenwich: JAI.

Lachmann, L. 1973. *Macro-economic thinking and the market economy*. London: Institute of Economic Affairs.

Lachmann, L. 1976. From Mises to Shackle: An essay on Austrian economics and the Kaleidic Society. *Journal of Economic Literature* 14: 54–62.

Lachmann, L. 1977. Austrian economics in the present crisis of economic thought. In *Capital, expectations, and the market process*. Kansas City: Sheed, Andrews & McMeel.

Lachmann, L. 1986a. Austrian economics under fire: The Hayek-Sraffa duel in retrospect. In Grassl and Smith (1986).

Lachmann, L. 1986b. *The market as a process*. Oxford: Basic Blackwell.
Lavoie, D. 1985. *Rivalry and central planning: The socialist calculation debate reconsidered*. Cambridge: Cambridge University Press.
Leser, N., ed. 1986. *Die Wiener Schule der Nationalökonomie*. Vienna: Hermann Böhlau.
Machlup, F. 1981. Ludwig von Mises: The academic scholar who would not compromise. *Wirtschaftspolitischen Blätter* 28 (4): 6–14.
Machlup, F. 1982. Austrian economics. In *Encyclopedia of economics*, ed. D. Greenwald. New York: McGraw-hill.
Mataja, V. 1884. *Der Unternehmergewinn.* Vienna.
Mayer, H. 1932. Der Erkenntniswert der Funktionellen Preistheorien. In *Die Wirtschaftstheorie der Gegenwart*, ed. H. Mayer. Vienna: Springer.
McCulloch, J.H. 1977. The Austrian theory of the marginal use and of ordinal marginal utility. *Zeitschrift für Nationalökonomie* 37 (3–4): 249–280.
Menger, C. 1871. *Grundsätze der Volkwirthschaftslehre.* Trans. as *Principles of economics,* ed. J. Dingwall and B.F. Hoselitz, 1950; reprinted. New York: New York University Press, 1981.
Menger, C. 1883. *Untersuchungen über die Methode der Sozialwissenschaften und der politischen Ökonomie insbesondere.* In Trans. F.J. Nock as *Problems of economics and sociology,* ed. L. Schneider, Urbana: University of Illinois Press, 1963. Reprinted as *Investigations into the method of the social sciences with special reference to economics* with a new Introduction by L.H. White, New York/London: New York University Press, 1985.
Menger, K. Jr. 1973. Austrian marginalism and mathematical economics. In *Carl Menger and the Austrian school of economics*, ed. J.R. Hicks and W. Weber. Oxford: Clarendon Press.
Meyer, R. 1887. *Das Wesen des Einkommens: Eine volkswirthschaftliche Untersuchung*. Berlin: Hertz.
O'Driscoll, G.P. Jr. 1977. *Economics as a coordination problem: The contributions of Friedrich A. Hayek*. Kansas City: Sheed, Andrews & McMeel.
O'Driscoll, G.P. Jr., and M.J. Rizzo. 1985. *The economics of time and ignorance*. Oxford: Basil Blackwell.
Orosel, G.O. 1981. Faber's modern Austrian capital theory: A critical survey. *Zeitschrift für Nationalökonomie* 41: 141–155.
Robbins, L. 1932. *The nature and significance of economic science*. London: Macmillan.
Rothbard, M.N. 1962. *Man, economy, and state: A treatise on economic principles*. Princeton: Van Nostrand.
Sax, E. 1887. *Grundlegung der Theoretischen Staatswirtschaft*. Vienna: Holder.
Schumpeter, J.A. 1908. *Das Wesen und der Hauptinhalt der Theoretischen Nationalökonomie*. Leipzig: Duncker & Humblot.
Schumpeter, J.A. 1912. *Theorie der wirtschaftlichen Entwicklung.* Leipzig: Duncker & Humblot. English trans. as *The theory of economic development*. Cambridge, MA: Harvard University Press, 1934.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.
Streissler, E. 1969. Structural economic thought: On the significance of the Austrian school today. *Zeitschrift für Nationalökonomie* 29 (3–4): 237–266.
Streissler, E. 1972. To what extent was the Austrian school marginalist? *History of Political Economy* 4: 426–461.
Streissler, E. 1973. The Mengerian tradition. In *Carl Menger and the Austrian school of economics*, ed. J.R. Hicks and W. Weber. Oxford: Clarendon Press.
Streissler, E. 1986. Arma virumque cano. Friedrich von Wieser, the bard as economist. In Leser (1986).
von Böhm-Bawerk, E. 1884. *Geschichte und Kritik der Kapitalzins-Theorien.* English trans. as vol. 1 of *Capital and interest.* South Holland: Libertarian Press, 1959.
von Böhm-Bawerk, E. 1886. Grundzuge der Theorie des Wirtschaftlichen Guterwerths. *Conrad's Jahrbuch* 13 (1–88): 477–541.
von Böhm-Bawerk, E. 1889. *Positive Theorie des Kapitales*. Innsbruck: Wagner.
von Böhm-Bawerk, E. 1891. The Austrian economists. *The Annals of the American Academy of Political and Social Science* 1: 361–384.
von Böhm-Bawerk, E. 1896. *Zum Abschluss des Marxschen Systems.* In Trans. (1898) as *Karl Marx and the close of his system,* ed. P. Sweezy. New York: Kelley, 1949.
von Komorzynski, J. 1889. *Der Werth in der isolirten Wirthschaft*. Vienna: Manz.
von Mises, L. 1912. *Theorie des Geldes und der Umlaufsmittel.* Trans. as *Theory of money and credit* (1934). Indianapolis: Liberty Classics, 1980.
von Mises, L. 1933. *Grundprobleme der Nationalökonomie.* Trans. as *Epistemological problems of economics.* Princeton: Van Nostrand, 1960.
von Mises, L. 1940. *Nationalökonomie Theorie des Handelns und Wirtschaftens*. Geneva: Editions Union.
von Mises, L. 1943. 'Elastic expectations' and the Austrian theory of the trade cycle. *Economica* 10: 251–252.
von Mises, L. 1949. *Human action: A treatise on economics*. New Haven: Yale University Press.
von Mises, L. 1969. *The historical setting of the Austrian school of economics*. New Rochelle: Arlington House.
von Mises, L. 1978. *Notes and recollections*. South Holland: Libertarian Press.
von Philippovich, E. Philippsberg. 1893. *Grundriss der Politischen Ökonomie*. Freiburg: Mohr.
von Wieser, F. 1884. *Ursprung des Wirtschaftlichen Wertes*. Vienna: Hölder.
von Wieser, F. 1914. *Theorie der Gesellschaftlichen Wirtschaft.* Tübingen: Mohr. Trans. as *Social economics*. London: G. Allen & Unwin, 1927; reprinted, New York: Kelley, 1967.
Walras, L. 1874. *Eléments d'économie politique pure*. Lausanne: Corbaz.

White, L.H. 1977. *The methodology of the Austrian school economists.* Revised ed. Auburn: The Ludwig von Mises Institute of Auburn University, 1984.
Wieser, F. 1889. *Der Naturliche Werth.* Vienna: Hölder. In Trans. as *Natural value,* ed. W. Smart. London: Macmillan, 1893; reprinted, New York: Kelley, 1956.
Zuckerkandl, R. 1889. *Zur Theorie des Preises*. Leipzig: Stein.

Austrian Economics: Recent Work

Mario J. Rizzo

Abstract

This article reviews research in Austrian economics over the last 25 years, relating it to (but not discussing in detail) earlier classic work in the Austrian tradition. Core issues are business cycle theory, entrepreneurship, market processes and economic institutions, the communication of knowledge in markets, spontaneous order, and issues related to law and economics.

Keywords

Austrian business cycle theory; Austrian economics; Entrepreneurship; Hayek, F.A; Law and economics; Market processes; Monetary theory; Spontaneous order; Time

JEL Classifications

B53

Introduction

In the past 25 years, a large amount of new research in Austrian economics has developed and expanded the basic themes that are central to its unique identity (O'Driscoll and Rizzo 1996). These highly interrelated themes are (1) the subjective, yet socially embedded, quality of human decision making; (2) the individual's perception of the passage of time ('real time'); (3) the radical uncertainty of expectations; (4) the decentralization of explicit and tacit knowledge in society; (5) the dynamic market processes generated by individual action, especially entrepreneurship; (6) the function of the price system in transmitting knowledge; (7) the supplementary role of cultural norms and other cultural products ('institutions') in conveying knowledge; and (8) the spontaneous – that is, not centrally directed – evolution of social institutions. The specific ways in which these themes have recently manifested themselves is the subject of this article.

Since our task is to discuss the developments in Austrian economics primarily since the last *New Palgrave* entry (1987) we shall not review the work of the many 'classic' Austrian authors. In addition, since our concerns are the substantive developments in the field, we omit many valuable contributions in the history of economic thought and in methodology.

Macroeconomics and Monetary Theory

There have been many advances in Austrian macroeconomics. These include new work on business cycle theory and on alternative monetary institutions.

Each of these areas can be looked at from the general perspective of treating time, money and their related institutions seriously (Horwitz 2000). Time is the medium of all action. Decisions are taken in time to produce consequences in the future. Taking time seriously means also taking the uncertainty that characterizes these decisions seriously. This applies to savings-investment choices, production plans, and the time structure of capital goods. In an Austrian (and Keynesian) perspective the pervasive uncertainty of the future makes money necessary. Thus, as time is the medium of all action, money is the medium of all exchange. All goods markets are accordingly affected by the supply and demand for money and the nature of monetary institutions.

The Austrian Business Cycle Theory (ABCT) received a major systemization and refinement in the work of Roger Garrison, culminating in his book, *Time and Money: The Macroeconomics of*

Capital Structure (2001). The previous work in the subject was scattered in many articles by Friedrich Hayek and in the work of Ludwig von Mises. It was also very imperfectly linked to the brilliant, but underrated, work by Ludwig Lachmann, *Capital and its Structure* (1956). Garrison corrects these deficiencies and adds coherence to ABCT which had previously been unknown. In a sense, Garrison has done for ABCT what John Hicks and Alvin Hansen did for Keynes's macroeconomics, except that the Garrison's work is an accurate rendition of Hayek, Mises and Lachmann.

The subtitle of Garrison's book, 'The Macroeconomics of Capital Structure', expresses the important claim that Austrian macroeconomics cannot adequately be appreciated without understanding that 'investment' is not a homogeneous decision. This insight is developed at length by Peter Lewin in *Capital in Disequilibrium: The Role of Capital in a Changing World* (1999), the most important work in Austrian capital theory in many decades (see also Endres and Harper 2008). The ABCT focuses on the inappropriateness of the capital structure (malinvestment) generated by artificially low real interest rates (that is, interest rates that are lower than the real supply of savings would allow). Thus, the term over-investment is, by itself, a misleading characterization of the ABCT process. While excessively low interest rates do increase the level of investment relative to its previous position, they do so in a biased way – those stages of production further from consumption are affected to a greater extent.

However, as Garrison's recent work (2004) has shown, there are even more widespread distortions in the production structure generated by artificially low interest rates. These include initial 'overconsumption' as the result of reduced savings and of increased incomes on the part of factors of production. Increased investment in close temporal proximity to the overconsumption is labeled the 'derived demand effect'. This is in addition to the 'discount effect', described above, which increases the profitability of new investment distant from consumption. These two contrary effects come at the expense of intermediate stages of production as well as reduced maintenance of existing capital at all stages. They may even result from the utilization of unused resources during periods of less than full employment. These effects show that the ABCT is a type of 'coordinationist macroeconomics' insofar as it describes the discoordination of various sectors of the economy, and is not simply a micro choice-theoretic approach to macroeconomics (Wagner 2005).

Accordingly, in this Austrian view recessions are characterized not simply by low levels of aggregate economic activity but also by the misdirection of resources caused by previous boom-induced malinvestments. These systematic sectoral imbalances – too much investment in interest-sensitive areas of economic activity – must be corrected as recovery proceeds.

The Austrian theory, however, is not a complete theory of the business cycle. It accounts mainly for the process leading to and including the cycle's upper turning point. It is a theory of the crisis. How long the resulting recession lasts is not predicted by the theory or even, strictly speaking, by the degree to which resources were misallocated. The length of the recession will depend, for example, on those factors affecting the mobility of resources.

None of this implies that Hayek, Garrison or Horwitz are insensitive to the problems that would be induced by an aggregate increase in the demand to hold money (a fall in income velocity), which can accompany recessions. This 'secondary deflation' should be avoided by a concomitant increase in the supply of money by the relevant monetary institutions. Horwitz (2000) is the first to integrate Austrian macroeconomics with monetary disequilibrium theory to analyse deflationary processes. Nevertheless, recessions are not primarily deflationary phenomena (or at least need not be), but occasions for correction of the misdirection of resources. Some Austrians, however, argue that increases in the demand for money have significant negative consequences only in the presence of legal restraints on price flexibility (Salerno 2003).

One of the most important possible obstacles to recovery from recessions may be in the behaviour of 'big players'. These are agents whose discretionary behaviour, insulated from the normal

discipline of profit and loss, can significantly affect the course of economic effects (Koppl and Mramor 2003; Koppl 2002; Koppl and Yeager 1996). Thus, discretionary behaviour on the part of monetary authorities (in the United States, the Fed), fiscal policy makers (Congress or the Executive), or even in some cases private monopolists, can increase uncertainty faced by most economic agents ('small players'). They will have to pay more attention to trying to guess the perhaps idiosyncratic behaviour of the big players. Economic variables will become contaminated with big-player influence. It will become more difficult to extract knowledge of fundamentals from actual market prices. And thus entrepreneurs will find it harder to determine where resources should be withdrawn and where they should be added in a way that is sustainable in the medium to long term.

An important variant of the ABCT in *Risk and Business Cycles: New and Old Austrian Perspectives* (1998), developed by Tyler Cowen, focuses on the integration of business cycle theory with developments in modern finance. The main sense in which this can be called a variant of ABCT is that changes in the riskiness of investment decisions are linked to the 'old Austrian' concern with the degree of futurity or roundaboutness in investments. For example, in Cowen's analysis, an increase in the acceptable level of risk will encourage undertaking more longer-term investments (as well as, of course, investments of any given length with more uncertain yields). These can be both investments in durable capital goods (that is, investments with a continuous flow of payoffs over a long period of time) and investments with a long period of gestation before the ultimate output is produced. Cowen associates less risky ('safe') investments with consumption and shorter-term investments.

Cowen's analysis is more general than the traditional ABCT because it allows many factors besides a fall in real interest rates to generate a lengthening of the capital structure. These include exogenous risk-preference shifts, increases in savings, easing financial constraints, and reductions in uncertainty (so as to reduce 'waiting' for acceptable investment opportunities). Any of these changes can generate an increase in the riskiness of investment. None of these changes must necessarily cause a cyclical boom and bust, but they might do so.

Horwitz (2000) shows that the traditional business cycle concerns of Austrian macroeconomics quite naturally lead into comparative institutional analysis. Therefore, the obvious question is: What kind of institutional framework is necessary or conducive to avoiding the distortionary effects of inflation and deflation? Austrians have been critical of both discretionary central banking policies and rigid monetarist rules. Some have favoured free banking while others have favoured a 100 per cent (usually gold) reserve requirement and hence have opposed fractional reserve banking.

The free banking school, represented by Selgin and White (1994), Horwitz (2000), Dowd (1996) and Sechrest (1993), emphasizes the importance of adjusting to changes in the demand to hold money (income velocity). For prices in particular markets to do their work appropriately in transmitting knowledge and allocating resources they must be free of the distortions induced by inflation and monetarily induced deflation. Free-banking advocates argue that bank profit maximization, under sound institutional constraints, will lead banks to expand or contract deposits or currency *pari passu* with changes in the demand for money. Banks will receive signals about the demand for (their) money as their reserves expand or contract. When reserves expand, the demand to hold is increasing, and vice versa. Profit maximization leads banks to increase the supply of money when reserves expand beyond their desired levels. Thus, no explicit monetary policy is needed to avoid unwarranted expansion or contraction on the 'money market', just as on commodity markets no deliberate industrial policy is needed to avoided unwarranted expansion or contraction of resources in different areas.

The advocates of 100 per cent reserve money follow the work of Murray N. Rothbard (2008). These include Block (1988), Hoppe (1994), and Huerta de Soto (1995). They argue that free banking – to the extent that it is fractional reserve banking – is ethically suspect. Regardless of the merits of this argument, our concern here is solely with economics. They further argue that fractional

reserve banks are inherently inflationary because any creation of fiduciary media beyond an increase in specie will generate a business cycle. (The word 'inflationary' is being used here either as a definition – an increase in money not covered by an increase in specie – or as an intellectual place-card to suggest the generation of a cycle.) Critiques are offered in Horwitz (2000) and in Selgin and White (1994).

Entrepreneurship

The theory of entrepreneurship has been a subject of great importance in Austrian economics since the publication of Israel Kirzner's *Competition and Entrepreneurship* (1973). One could argue, of course, that this was implicit in the prior works of Ludwig von Mises and Friedrich Hayek. Nevertheless, there is an important difference between implicit and explicit ideas. Over a long period of time, Kirzner refined his theory of entrepreneurial discovery or alertness in many books.

Kirzner's approach is predominantly cognitive. There are roots of this cognitive approach in the early work of von Mises (Ebeling 2007). However, quite curiously in this time of the resurgence of psychology and economics, it is a cognitive theory without explicit cognitive foundations. Kirzner is interested in the market implications of the fact that there is entrepreneurial alertness. He is not interested, beyond some very general observations, in the causal factors that give rise to or are conducive to alertness.

Alertness, or equivalently, entrepreneurial discovery, is hard to define. It is a creative, spontaneous, and to a certain extent idiosyncratic, mental act that goes beyond the mere apprehension of objective data. First, while it usually begins with objective data, it critically involves drawing connections with other data when those connections are not obvious or even the result of complex computations. Second, true discoveries are not the result of deliberate acts of search. They cannot reliably be attained by the simple deployment of resources. Something more is necessary. This is not to suggest, however, that they must be viewed as random shocks to the economic system. They can be cultivated and prepared for by deliberate decisions, but they cannot be mechanically produced by them. We might say that while deliberate search is not a sufficient condition for discovery, it is necessary. Even better, eschewing the excessively constraining categories of necessity and sufficiency, we might say that the serendipity of discovery favours the searching mind (Holcombe 2007; Shane 2000). Finally, it is likely that many individuals can be exposed to the same data and yet not make the discoveries that the alert individual does.

In a market context, entrepreneurial cognition is the discovery of profit opportunities. In Kirzner's perspective, this is based on noticing price inconsistencies, whether at a point in time or across time. Hence this is an arbitrage theory of profit. How well this conception of entrepreneurship takes uncertainty into account is a matter of some dispute (Kirzner 1982). In the theory advanced by Young Back Choi (1993, 1999), however, uncertainty is more explicitly considered. In this perspective, related to Schumpeter's classic analysis (1934), entrepreneurs break through the conventional way of looking at the world. These conventions were originally adopted to reduce uncertainty. But as time goes on the world changes and they become less and less effective. Profit opportunities accumulate. Entrepreneurs adopt new paradigms that enable them to see the new profit opportunities that conventionalists cannot.

To the extent that entrepreneurial discovery is unconnected to any cognitive or psychological basis, it functions as a *deus ex machina* of the market process. It drives the processes that occur in response to errors and disequilibria. Ultimately, it is defined by what it does. This approach has been criticized because it presupposes empirical psychological processes that are not necessarily present in all circumstances (Jakee and Spong 2003). Can we say anything systematic about the factors that, on an individual or social level, are conducive to discovery? If we can, we might begin to understand more precisely what it is, when it is successful and when it is not.

In the first unified analysis of the factors affecting entrepreneurship, David Harper focuses

on the presence of a sense of personal agency as the primary factor. 'It comprises two cognitive elements – beliefs in the locus of control (or contingency expectations) and beliefs in self-efficacy (or competence expectations)' (Harper 2003, p. 14). This means that the entrepreneurial agent believes that in a particular context results are contingent upon actions as opposed to luck or nature, and that he himself possesses the personal capabilities to effect these actions and thus to produce the overall results. Individual characteristics also interact with situations to make the development of a discovery propensity more likely. Harper goes on to show the ways in which economic, political and cultural institutions mediate the individual factors.

In most Austrian treatments entrepreneurial discovery is important because it drives the market process. Nevertheless, in path-breaking work Frederic Sautet (2000) shows that there are multiple levels of entrepreneurship. In the simple case, the entrepreneur is herself alert to profit opportunities outside of the firm. In the more complex case, the entrepreneur must face the fact that she often doesn't know what her employees know. They are often closer to the local facts and may have a superior insight in some respects about profit opportunities in the firm (from restructuring) as well as outside of the firm. Thus, the entrepreneur in a 'complex firm' will seek to structure the firm with abstract or loose rules – some relating to compensation schemes – that encourage employees to make discoveries and communicate those appropriately. The firm itself can be a locus of entrepreneurship. In related work, Harper (2008b) suggests that a team of individuals, either inside or outside firms, might also constitute an entrepreneurial unit.

Randall Holcombe (2007) utilizes an idea of entrepreneurship beyond pure cognitive alertness, which includes, as well, acting upon the perception of novel opportunities. In this view, the entrepreneur can never be certain that she has correctly perceived a profit opportunity until she acts and assesses the consequences.

Some Austrians have not followed Kirzner in their analysis of entrepreneurship. For example, Joseph Salerno (1993) rejects the characterization of alertness as the essence of entrepreneurship. He sees resource ownership as a necessary feature of entrepreneurial activity (Salerno 2008). Along similar lines, K. Foss, N. Foss and P. Klein (2007, see also Klein 2008; Foss and Foss 2007) have weaved together aspects of Knight's uncertainty theory (1971) and Austrian heterogeneous capital theory (Lachmann 1956) to create a theory of entrepreneurial judgment. This theory makes entrepreneurship inseparable from asset ownership. The entrepreneur's judgement is about the control of heterogeneous capital assets under conditions of radical uncertainty. These authors have applied their theory to understanding the internal operation of the firm.

Market Processes and Economic Institutions

The entrepreneurial function is closely related to market processes and economic institutions. These interrelations are both complex and important. It will help to somewhat artificially separate them for our consideration.

A. The Austrian approach to market processes is distinctive in a number of respects (Wagner 2007, 2010). It is sometimes described as a genetic-causal theory (Cowan and Rizzo 1996). First, markets are in process and not continually in equilibrium. Thus, most Austrians do not take interpersonal equilibria of any kind simply as given or as consequences of an axiom of rationality. (An exception is Salerno 1994, who considers momentary market-clearing equilibrium as an implication of rationality.) Lack of alertness can be responsible for economic errors and inconsistencies (or lack of interpersonal coordination). The market process consists of those entrepreneurial responses to error. Kirzner and others take the view that market processes are generally coordinating: that is, that they generally correct market errors. Austrians accept this as an *empirical generalization*. The extent to which the empirical generalization can be traced to an a priori discovery tendency is a subject of debate.

Kirzner appears to accept this view because he sees the tendency to discover as equivalent to, or tightly connected to, the tendency toward greater coordination (Kirzner 1997). Rizzo, however, rejects this equivalence (Rizzo 1996). Other authors have also expressed similar, though not identical, criticisms (Klein and Briggerman 2009; Klein 1997). The neoclassical view that equilibrium is an implication of rationality should not, in this author's opinion, be replaced with the view that a tendency to equilibrium is the implication of purposefulness. The former has empirical implications while the latter is not clearly defined, unless it is meant as the positive heuristic of an empirical research program (Rizzo 1982).

Second, market processes are not instantaneous but take time. In the passage of time ('real time'), knowledge changes and unpredictable events occur (O'Driscoll and Rizzo 1996). What were data at the start of a process may change because the process of 'equilibration' occurs in real time. Real time cannot elapse without knowledge changing.

Third, market processes take place in the context of radical uncertainty. This is to be distinguished from risk, in which all of the possibilities are known with objective probabilities. However, radical uncertainty is not simply a condition where the assigned probabilities are not objective, but one in which not all of the possibilities are known beforehand. (Still further complications ensue because sometimes individuals know that they don't know the possibilities and sometimes they do not.)

This leads to the fourth feature of market processes: they are relatively indeterminate. If market processes – in the form of entrepreneurial discovery – cannot be predicted, then the economist cannot know at the beginning where they will lead. In the process of adjusting to change, new 'data' will be discovered (Rizzo 2000; 1990). How far to take this point about the indeterminacy of market processes is subject to debate and may, in part, depend on definitional issues (Holcombe 2007). Some have argued that Kirzner, in particular, has incorrectly downplayed this indeterminacy (Jakee and Spong 2003).

This is not to rule out the use of constructs in which equilibria are reached as heuristic devices when appropriate (Holcombe 2007). However, since they are simply heuristic devices they can be thrown out when circumstances do not warrant such 'static' dynamics.

The fifth, and final, feature of market processes is the communication of decentralized or scattered knowledge. Markets enable individuals to act on more knowledge than they can ever hope to possess explicitly. They can do this through entrepreneurially produced market prices and through non-price manifestations of market behaviour. As Hayek showed, the man on the spot may be directly aware of certain economically relevant conditions. If he acts by taking advantage of this knowledge in profitably buying or selling he will ensure that market prices communicate what he knows (Hayek 1948; Kirzner 1992a).

Prices are not the only communicators of knowledge in markets. Capital goods also embody knowledge. First, the particular use and combination of capital goods can, under non-distortionary conditions, convey knowledge about efficient resource allocation and possible profit opportunities (Lachmann 1956). Second, even the physical design of capital goods can convey accumulated knowledge about successful production techniques (Baetjer 2000).

In general, the communication of knowledge in market settings depends not only on catallactic phenomena but also crucially on the appropriate 'institutional' context. This includes legal and cultural products (Harper 2003). In the latter category David Harper (2008a, forthcoming) has drawn attention to the role of numerical cognition – a product of both unique human biology and cultural development – in facilitating economic calculation. The development of conventionalized systems of number sequences and techniques of counting reduces transaction costs, and helps agents to make plans, compute values, scarcities, notice arbitrage opportunities, and

ascertain the economically relevant aspects of capital goods.

B. Entrepreneurship does not simply operate within a familiar institutional structure like the market. It can also operate within structures like those involving social ties, philanthropy, non-profit organizations and so forth (Boettke and Coyne 2009). There is also 'political entrepreneurship' within a given constitutional or governance structure, which seeks to create coalitions to effect specific legislation or transfers of wealth ('rent seeking'). These non-market structures determine the precise form that entrepreneurship takes. The common differentiating factor that separates the entrepreneurship of the market process from these other forms of entrepreneurship is the absence of the discipline of monetary profit and loss in the latter cases. Although money may change hands as a result of these forms of entrepreneurial activity, their outputs are not valued according to market prices. Whether effective feedback mechanisms exist in these contexts is an open question (Boettke and Coyne 2009).

Some Austrians, however, have emphasized that non-market institutions can indeed provide feedback to entrepreneurs and can generate a social learning or knowledge-communication process similar to market prices and profit–loss signals (Chamlee-Wright 2008; Chamlee-Wright and Myers 2008; Lewis and Chamlee-Wright 2008). In particular, reputation and status are forms of 'social capital' that convey information. Under conditions of competition and effective monitoring of standards, knowledge can be transmitted far beyond networks of individuals in direct communication with each other.

An important example of the communication of knowledge in a non-market context can be found in the scientific community (McQuade and Butos 2003). We discuss this below in the section on spontaneous orders.

Entrepreneurship can also shape or create institutions. Rules of behaviour that surround and define markets, constitutional systems, social and cultural systems arise out of the previous framework of rules, whether it was de facto or de jure. (In fact, the distinction between de facto and de jure may not be all that important for the economics of institutions, aside from the possible issue of transaction costs.) There is path dependency in the development of institutions (Boettke et al. 2008). Those that develop as 'indigenously introduced endogenous institutions' are closely related to the informal practices and expectations of people, which in turn are grounded in local knowledge and values. Other institutions may be indigenously introduced but are exogenous in the sense that they are imposed by some formal authority, and do not gradually evolve from the informal traditions of a people. There is a risk that these institutions will not 'stick' because of conflict between the institution and the underlying norms. Externally (or foreign) introduced exogenous institutions exhibit the greatest probability of not succeeding because of the greater likelihood of conflict with underlying norms and expectations. Boettke et al. (2008) refer to this analysis as an example of the 'regression theorem' first propounded by Ludwig von Mises (1953) in his analysis of the evolution of money.

Some of the evolution of framework institutions may simply be the undesigned outcome of individual behaviour that is not necessarily entrepreneurial, as when people follow each other in making a path through the snow (Kirzner 1992b). In other cases, there may be alertness to possibilities of gain for the relevant acting parties in altering the political or social frameworks. Plausibly, the creation of the US Constitution was one such case.

Institutions exist at many levels. Perhaps the most basic are those that involve informal institutions like customs, traditions, norms and religion (Williamson 2000). These take the longest time to change. They may also determine the standards by which lower-level institutions and behaviour within them are evaluated. A new political system is good or bad depending on the (more basic) norm structure in place.

Spontaneous Orders

Our discussion of entrepreneurship and of institutions leads naturally into a discussion of

spontaneous order, an idea very closely associated with Austrian economics. Unfortunately, the term 'spontaneous order' is opaque. Somewhat more descriptive is the expression made famous by F.A. Hayek, 'the results of human action but not of human design' (Hayek 1967), and even more descriptive is the idea of unintended social order produced by individually purposeful behaviour.

A spontaneous order is an organic or emergent form of coordination that manifests itself in social institutions, some organizations and clusters of individual plans. Orders of this kind arise without the design and maintenance (oversight) of a social planner. Nevertheless, spontaneous orders are generated by individual agents who do plan and carry out actions within their sphere of activity. Social order emerges as individuals adjust their plans to each other and to the environment over time.

Spontaneous order theories come in different varieties. Some refer to order produced on markets, while others concern order produced in non-market settings. These theories can be purely positive (descriptive) or they can also be normative. When they are normative their normativity can be relative to the society as a whole or simply to particular subgroups.

At the most basic level, spontaneous order can refer simply to the welfareenhancing outcomes of competitive market processes operating within the 'fixed' constraints of property, contract and tort law. This is best studied within the context of market entrepreneurship.

Bruce Benson (1989), in his path-breaking study of the spontaneous evolution of commercial law, shows how market interactions, based on basic property constraints, can give rise to commercial (contract) law without a law-giver. The self-interested interactions of merchants lead them to develop and adhere to rules that increase their trade and hence overall social cooperation. These rules develop through a process of trial and error in which entrepreneurial alertness at a higher level – the level of rules of the game – doubtless plays an important role. Similarly, Stringham (2002, 2003) and Stringham and Boettke (2004) show that the selfinterested interaction of participants in financial markets has generated useful regulations that govern the operation of these markets.

Peter Leeson (2007, 2009), in a number of studies of the organization of 18th century pirate activity, shows how an outlaw subgroup of society developed maximizing (or 'rational' in a limited sense) rules of governance without central direction. Outside of a market context, pirates converged on a set of rules whereby their ability to steal wealth from the rest of society was enhanced. This involved rules within the pirate society itself as well as rules governing its treatment of others. Within their society 'democracy' was used; outside of it the use of brutality was constrained. This case is a good example of a spontaneous ordering process with 'good' consequences within the subgroup and yet negative consequences for society as a whole. Pirates steal resources from the rest of society. The success of any such rogue subgroup weakens the possibilities of voluntary exchange and other forms of peaceful interaction.

Thomas McQuade and William Butos (2003; see also Butos and Koppl 2003; Butos and McQuade 2006) further develop the spontaneous-order approach in the case of the organization of scientific research communities. Even where markets in the traditional sense may be missing, spontaneous – that is, non centrally directed – ordering processes are still present. They focus on the evolved non-market mechanism of publication–citation–reputation. Scientific knowledge is viewed as a 'by-product' of the intentional activities of scientists to publish their results, get citations and enhance their reputations. Within this process competition among scientists tends to filter out inferior ideas. The resultant product ('science') is orderly in the sense that it tends to be reliable and codifiable. A set of procedures is put into place which acts as a filter to discriminate between rival claims. Furthermore, what comes out of the filter can be collected, integrated with other knowledge and transferred to other scientists.

These illustrations suggest the need for a more general theory of spontaneous order that would clarify the various conditions under which such ordering-processes will take place. Specifically, it should also explore the role of markets and market prices, since it is clear that spontaneous order can develop without markets. From the welfare point

of view the research discussed above leaves us with a puzzle: When do spontaneous orders produce an enhancement of social welfare and when a reduction in it, as in the case of pirate societies?

Law and Economics

One of the most important areas of research in Austrian economics is the vibrant area of law and economics. Some of the contributions mentioned above in connection with spontaneous order and institutions could be included in this section. The field's uniquely Austrian features consist of attention to (1) the process of law and state intervention in markets; (2) the need for relatively stable law in a world of external change; (3) the influence of decentralized knowledge on the character and limits of law; and (4) the privatization of some of the basic functions of the state.

1. The most significant work on the processes generated by intervention since the classic analyses of Ludwig von Mises (1977), F.A. Hayek (1994) and Israel Kirzner (1985) can be found in Sanford Ikeda (1997, 2003, 2005) and in Mario Rizzo and Glen Whitman (2003). Ikeda's framework focuses on the deviation of the actual outcomes of intervention from the intended outcomes. This gap, based on an assumption of radical ignorance, generates price distortions, whether because the intervention takes the form of price regulations or because redistribution of wealth degrades incentives and thus individual responses to underlying economic data. These economic changes interact with largely, though not entirely, endogeneous changes in ideology to produce a tendency toward further policy intervention.

   Rizzo and Whitman, on the other hand, begin from the largely philosophical and jurisprudential literature of 'slippery slopes'. They construct a general approach that emphasizes the role of changes in ideas, or more precisely, in the arguments that rationalize or justify legislative policies or judicial decisions. The mechanism by which these arguments change is a combination of the largely unanticipated consequences of decisions and the higher-level theories in which acceptable arguments are embedded.

   Recently, Rizzo and Whitman (2009; see also Whitman and Rizzo 2007) have applied their slippery slope analysis in conjunction with many of the assumptions and findings of behavioral economics to demonstrate the expansive tendencies inherent in the supposedly moderate policies of new or 'libertarian' paternalism.

   The Rizzo–Whitman and Ikeda approaches seem largely compatible. Ikeda stresses more traditional economic processes, while Rizzo and Whitman stress the details of the intellectual changes that occur in the context of economic or other processes. In neither of these approaches is the 'slippery slope' consequence of policies inevitable. They each describe tendencies that could be counterbalanced in specific cases, but which often have not been.
2. The classic work of Hayek (1960, 1973) on the rule of law simultaneously stresses the importance of stability in the legal framework and its adaptability to changing external circumstances. The solution to this paradox can be found in the level of abstraction of the relevant rules. For example, the abstract form of contract law can remain stable while the prices, conditions and content of exchanges vary at a point in time or over time. The consequences of abstraction in legal rules are examined in Whitman (2009). Whitman shows that an intermediate level of abstraction is optimal from the perspective of generating rules with predictable consequences.

   From a slightly different perspective, Rizzo (1980a, b, 1985) and Roy Cordato (2007) both criticize the cost–benefit framework in many conceptions of negligence law because it produces legal decisions that lack predictability to those for whom the particular law is relevant. Peter Lewin (1982) extends the critique to pollution externalities and social cost. The economic data upon which efficient legal decisions are to be made are often unavailable, complex or transient. This is especially true in a world characterized by radical uncertainty.

Thus the so-called economic approach to tort law is defective on its own terms. Lack of predictability generates costs. In terms of the abstraction language of Whitman's analysis, the problem of the efficiency approach is that it enshrines a standard, rather than a set of specific rules, which is too abstract.

Similar criticisms of the so-called economic approach to property rights that derives from Ronald Coase and Harold Demsetz have been advanced by Walter Block. Block argues that the Coasian cost–benefit approach effectively abolishes property rights (1977 1977, 1995, 2000). This view is extended to the analysis of the recent US Supreme Court eminent domain case, Kelo *v.* City of New London (Block 2006).

3. The decentralization of factual knowledge is a critically important factor limiting the feasibility of many forms of intervention. As in the earlier analysis of Mises (1977), the critiques discussed here begin from the announced goals of the interveners and do not challenge their worthiness. The approach is thus non-normative. It simply seeks to answer the question: Can the policies achieve the goals that their advocates have set? Rizzo (2005) tackles this question in the case of moral paternalism: that is, the form of paternalism that coerces the individual in the interests of her moral betterment. Using the internal standards of three major ethical approaches – utilitarianism, natural law and Kantianism – Rizzo argues that the factual knowledge needed to determine just what the moral course of action is in concrete cases is not available to the paternalist. Rizzo and Whitman (forthcoming) also apply this kind of analysis to a form of economic paternalism based on behavioural economics. They argue that the factual knowledge that behavioural economics claims is relevant to the crafting of policies designed to improve the decisions of individuals exceeds what is known to the policy makers.
4. Most economic analysis proceeds on the assumption that the state exercises at least its minimum functions: that is, provision of protection, enforcement of property rights and contracts, and the adjudication of disputes. Nevertheless some economists in the broad Austrian and spontaneous order tradition have argued that privatization of at least some of these functions is feasible and desirable. Bryan Caplan and Edward Stringham (2008) have compared the private and public adjudication of disputes. They find that private adjudication is more efficient in areas of commercial disputes, and more generally in those areas where prior relationships exist among the parties. They also speculate on a broader use of private adjudication. In a related area of public choice economics, Powell and Stringham (2009) survey a surprisingly large extant literature on the economics of a stateless society.

See Also

- ▶ Austrian Economics: Recent Work
- ▶ Business Cycles
- ▶ Entrepreneurship
- ▶ Spontaneous Order

Bibliography

Baetjer, H. 2000. Capital as embodied knowledge: Some implications for the theory of economic growth. *Review of Austrian Economics* 13: 147–174.

Benson, B. 1989. The spontaneous evolution of commercial law. *Southern Economic Journal* 55: 644–661.

Block, W. 1977. Coase and Demsetz on private property rights. *Journal of Libertarian Studies* 1: 111–115.

Block, W. 1988. Fractional reserve banking. In *Man, economy and liberty: Essays in honor of Murray H. Rothbard*, ed. W. Block and L. Rockwell, 24–31. Auburn: Mises Institute.

Block, W. 1995. Ethics, efficiency, Coasian property rights, and psychic income: A reply to Demsetz. *Review of Austrian Economics* 8: 61–125.

Block, W. 2000. Private-property rights, erroneous interpretations, morality, and economics: Reply to Demsetz. *Quarterly Journal of Austrian Economics* 3: 63–78.

Block, W. 2006. Coase and Kelo: Ominous parallels and reply to Lott on Rothbard on Coase. *Whittier Law Review* 27: 997–1022.

Boettke, P., and C. Coyne. 2002. Entrepreneurship and development: Cause or consequence? *Advances in Austrian Economics* 6: 67–88.

Boettke, P., and C. Coyne. 2009. Context matters: Institutions and entrepreneurship. *Foundations and Trends in Entrepreneurship* 5: 135–209.

Boettke, P.J., C.J. Coyne, and P.T. Leeson. 2008. Institutional stickiness and the new development economics. *American Journal of Economics and Sociology* 67: 331–358.

Butos, W.N., and R. Koppl. 2003. Science as a spontaneous order. In *The evolution of scientific knowledge*, ed. H.S. Jensen, L.M. Richter, and M.T. Vendelø, 164–188. Northampton: Edward Elgar.

Butos, W.N., and T.J. McQuade. 2006. Government and science: A dangerous liaison? *Independent Review* 11: 177–208.

Caplan, B., and, E.P. Stringham. 2008. Privatizing the adjudication of disputes. *Theoretical Inquiries in Law* 9, Article 8. Available at: http://www.bepress.com/til/default/vol9/iss2/art8

Chamlee-Wright, E. 2008. The structure of social capital: An Austrian perspective on its nature and development. *Review of Political Economy* 20: 41–58.

Chamlee-Wright, E., and J.A. Myers. 2008. Discovery and social learning in nonpriced environments: An Austrian view of social network theory. *Review of Austrian Economics* 21: 151–166.

Choi, Y.B. 1993. *Paradigms and conventions: Uncertainty, decision making, and entrepreneurship*. Ann Arbor: University of Michigan Press.

Choi, Y.B. 1999. Conventions and learning: A perspective on the market process. In *Economic organisation and economic knowledge*, ed. S.C. Dow and P. Earl, 57–75. London: Edward Elgar.

Cordato, R. 2007. *Efficiency and externalities in an open-ended universe*. Auburn: Ludwig von Mises Institute, 1992.

Cowan, R., and M.J. Rizzo. 1996. The genetic-causal tradition and modern economic theory. *Kyklos* 49: 273–317.

Cowen, T. 1998. *Risk and business cycles: New and old Austrian perspectives*. London: Routledge.

Dowd, K. 1996. *Laissez-faire banking*. London: Routledge.

Dowd, K. 2000. *Money and the market: Essays on free banking*. London/New York: Routledge.

Ebeling, R.M. 2007. Austrian economics and the political economy of freedom. *New Perspectives on Political Economy* 3: 87–104. Available at: http://pcpe.libinst.cz/nppe/3_1/nppe3_1.pdf

Endres, A.M., and, D.A. Harper. 2008. *Capital as a layer cake: Menger, Lachmann and the nature of capital*. Unpublished manuscript, New York University.

Foss, K., and N.J. Foss. 2006. The limits to designed orders: Authority under 'distributed knowledge' conditions. *Review of Austrian Economics* 19: 261–274.

Foss, K., and N.J. Foss. 2007. The entrepreneurial organization of heterogeneous capital. *Journal of Management Studies* 44: 1165–1186.

Foss, K., N.J. Foss, and P.G. Klein. 2007. Original and derived judgment: An entrepreneurial theory of economic organization. *Organization Studies* 28: 1–20.

Garrison, R. 2001. *Time and money: The macroeconomics of capital structure*. London: Routledge.

Harper, D.A. 1998. Institutional conditions for entrepreneurship. *Advances in Austrian Economics* 5: 241–275.

Harper, D.A. 2003. *Foundations of entrepreneurship and economic development*. London: Routledge.

Harper, D.A. 2008a. A bioeconomic study of numeracy and economic calculation. *Journal of Bioeconomics* 10: 101–126.

Harper, D.A. 2008b. Towards a theory of entrepreneurial teams. *Journal of Business Venturing* 23: 613–626.

Harper, D.A. Forthcoming. Numbers as a cognitive and social technology: An economic ontology. *Journal of Institutional Economics*.

Hayek, F.A. 1944. *The road to serfdom*. Chicago: University of Chicago Press, 1994.

Hayek, F.A. 1948. The use of knowledge in society. In *Individualism and economic order*. Chicago: University of Chicago Press.

Hayek, F.A. 1960. *The constitution of liberty*. Chicago: University of Chicago Press.

Hayek, F.A. 1967. The results of human action but not of human design. In *Studies in philosophy, politics and economics*. London: Routledge & Kegan Paul.

Hayek, F.A. 1973. *Law, legislation and liberty, Vol. 1: Rules and order*. Chicago: University of Chicago Press.

Holcombe, R.G. 2007. *Entrepreneurship and economic progress*. London: Routledge.

Hoppe, H.-H. 1994. How is fiat money possible? or, the devolution of money and credit. *Review of Austrian Economics* 7: 49–74.

Horwitz, S. 2000. *Microfoundations and macroeconomics: An Austrian perspective*. London: Routledge.

Huerta de Soto, J. 1995. A critical analysis of central banks and fractional-reserve free banking from the Austrian perspective. *Review of Austrian Economics* 8: 25–38.

Huerta de Soto, J. 2006. *Money, bank credit, and economic cycles*. Trans. M.A. Stroup. Auburn: Ludwig von Mises Institute.

Ikeda, S. 1997. *Dynamics of the mixed economy: Toward a theory of interventionism*. London: Routledge.

Ikeda, S. 2003. How compatible are public choice and Austrian political economy? *Review of Austrian Economics* 16: 63–75.

Ikeda, S. 2005. The dynamics of intervention: Regulation and redistribution in the mixed economy. *Advances in Austrian Economics* 8: 21–57.

Jakee, K., and H. Spong. 2003. Praxeology, entrepreneurship and the market process: A review of Kirzner's contribution. *Journal of the History of Economic Thought* 25: 461–486.

Kirzner, I.M. 1973. *Competition and entrepreneurship*. Chicago: University of Chicago Press.

Kirzner, I.M. 1979. *Perception, opportunity, and profit: Studies in the theory of entrepreneurship*. Chicago: University of Chicago Press.

Kirzner, I.M. 1982. Uncertainty, discovery, and human action: A study of the entrepreneurial profile in the

Misesian system. In *Method, process, and Austrian economics*. Lexington: Lexington Books.
Kirzner, I.M. 1985. The perils of regulation: A market process approach. In *Discovery and the capitalist process*. Chicago: University of Chicago Press.
Kirzner, I.M. 1992a. Prices, the communication of knowledge and the discovery process. In *The meaning of market process*. London: Routledge.
Kirzner, I.M. 1992b. Knowledge problems and their solutions: Some relevant distinctions. In *The meaning of market process*. London: Routledge.
Kirzner, I.M. 1997. Entrepreneurial discovery and the competitive market process: An Austrian approach. *Journal of Economic Literature* 35: 60–85.
Klein, D.B. 1997. Convention, social order, and the two coordinations. *Constitutional Political Economy* 8: 319–335.
Klein, P.G. 2008. Opportunity discovery, entrepreneurial action, and economic organization. *Strategic Entrepreneurship Journal* 2: 175–190.
Klein, D.B., and J. Briggerman. 2009. Israel Kirzner on coordination and discovery. *Journal of Private Enterprise* 25: 1–53.
Knight, F.H. 1921. *Risk, uncertainty, and profit*. Chicago: University of Chicago Press, 1971.
Koppl, R. 2002. *Big players and the economic theory of expectations*. Hampshire/New York: Palgrave Macmillan.
Koppl, R., and D. Mramor. 2003. Big players in Slovenia. *Review of Austrian Economics* 16: 253–269.
Koppl, R., and L.B. Yeager. 1996. Big players and herding in asset markets: The case of the Russian ruble. *Explorations in Economic History* 33: 367–383.
Lachmann, L.M. 1956. *Capital and its structure*. London: G. Bell & Sons.
Leeson, P. 2007. An-arrgh-chy: The law and economics of pirate organization. *Journal of Political Economy* 115: 1049–1094.
Leeson, P.T. 2009. *The invisible hook: The hidden economics of pirates*. Princeton: Princeton University Press.
Lewin, P. 1982. Pollution externalities: Social cost and strict liability. *Cato Journal* 2: 205–229.
Lewin, P. 1999. *Capital in disequilibrium: The role of capital in a changing world*. London: Routledge.
Lewis, P., and E. Chamlee-Wright. 2008. Social embeddedness, social capital and the market process: An introduction to the special issue on Austrian economics, economic sociology and social capital. *Review of Austrian Economic s* 21: 107–118.
McQuade, T.J., and W.N. Butos. 2003. Order-dependent knowledge and the economics of science. *Review of Austrian Economics* 16: 133–152.
von Mises, L. 1912. *The theory of money and credit*. New Haven: Yale University Press, 1953.
von Mises, L. 1929. *A critique of interventionism*. Trans. H.F. Sennhollz. New Rochelle: Arlington House, 1977.
O'Driscoll Jr., G.P., and M.J. Rizzo. 1996. *The economics of time and ignorance*. London: Routledge.
Powell, B.W., and E.P. Stringham. 2009. Public choice and the economic analysis of anarchy: A survey. *Public Choice* 140: 503–538.
Rizzo, M.J. 1980a. Law amid flux: The economics of negligence and strict liability in tort. *Journal of Legal Studies* 9: 291–318.
Rizzo, M.J. 1980b. The mirage of efficiency. *Hofstra Law Review* 8: 641–658.
Rizzo, M.J. 1982. Mises and Lakatos: A reformulation of Austrian methodology. In *Method, process, and Austrian economics*, ed. I. Kirzner. Lexington: Lexington Books.
Rizzo, M.J. 1985. Rules versus cost–benefit analysis in the common law. *Cato Journal* 4: 865–884.
Rizzo, M.J. 1990. Hayek's four tendencies towards equilibrium. *Cultural Dynamics* 3: 12–31.
Rizzo, M.J. 1996. Introduction: Time and ignorance after ten years. In *The economics of time and ignorance*, ed. G.P. O'Driscoll Jr. and M.J. Rizzo. London: Routledge.
Rizzo, M.J. 2000. Real time and relative indeterminacy in economic theory. In *Time in contemporary intellectual thought*, ed. P. Baert, 171–188. Amsterdam: Elsevier.
Rizzo, M.J. 2005. The problem of moral dirigisme: A new argument against moralistic legislation. *NYU Journal of Law and Liberty* 1: 789–843.
Rizzo, M.J., and D.G. Whitman. 2003. The camel's nose is in the tent: Rules, theories, and slippery slopes. *UCLA Law Review* 51: 539–592.
Rizzo, M.J., and D.G. Whitman. 2009a. Little brother is watching you: New paternalism on the slippery slopes. *Arizona Law Review* 51: 685–739.
Rizzo, M.J., and D.G. Whitman. 2009b. The knowledge problem of new paternalism. *Brigham Young University Law Review* 2009: 905–968.
Rothbard, M.N. 1983. *The mystery of banking*. Auburn: Ludwig von Mises Institute, 2008.
Salerno, J.T. 1993. Mises and Hayek dehomogenized. *Review of Austrian Economics* 6: 113–146.
Salerno, J.T. 2003. An Austrian taxonomy of deflation – With applications to the U. S. *Quarterly Journal of Austrian Economics* 6: 81–109.
Salerno, J.T. 2008. The entrepreneur: Real and imagined. *Quarterly Journal of Austrian Economics* 11: 188–207.
Sautet, F.E. 2000. *An entrepreneurial theory of the firm*. London: Routledge.
Schumpeter, J.A. 1911. *Theory of economic development*. Cambridge, MA: Harvard University Press, 1934.
Sechrest, L. 1993. *Free banking: Theory, history, and a Laissez-Faire model*. Westport: Quorum.
Selgin, G.A., and L.H. White. 1994. How would the invisible hand handle money? *Journal of Economic Literature* 32: 1718–1749.
Shane, S. 2000. Prior knowledge and the discovery of entrepreneurial opportunities. *Organization Science* 11: 448–469.
Stringham, E. 2002. The emergence of the London stock exchange as a selfpolicing club. *Journal of Private Enterprise* 17: 1–19.

Stringham, E. 2003. The extralegal development of securities trading in seventeenth century Amsterdam. *Quarterly Review of Economics and Finance* 43: 321–344.
Stringham, E., and P. Boettke. 2004. Brokers, bureaucrats and the emergence of financial markets. *Managerial Finance* 30: 57–71.
Wagner, R. 2005. Austrian cycle theory and the prospect of a coordinationist macroeconomics. In *Modern applications of Austrian thought*, ed. J.G. Backhaus, 77–92. London/New York: Routledge.
Wagner, R. 2007. Value and exchange: Two windows for economic theorizing. *Review of Austrian Economics* 20: 57–68.
Wagner, R. 2010. *Mind, society, and human action: Time and knowledge in social economy*. London/New York: Routledge.
Whitman, D.G., and M.J. Rizzo. 2007. Paternalist slopes. *NYU Journal of Law and Liberty* 2: 411–443.
Whitman, D.G. 2009. The rules of abstraction. *Review of Austrian Economics* 22: 21–42.
Williamson, O.E. 2000. The new institutional economics: Taking stock, looking ahead. *Journal of Economic Literature* 38: 595–613.

Autarky

David Evans

Autarky means self-sufficiency, especially economic self-sufficiency. The term appears most frequently in economic literature, both as a theoretical construct deployed in the theory of comparative advantage, and as a policy of economic self-sufficiency.

In the theory of comparative advantage, the concept of autarky plays a central role. Originally developed by Torrens and Ricardo, the theory proceeds by considering at least two hypothetical commodity-producing economies. Each economy is supposed to be capable of existing in at least two states, one of which is autarky or no trade, and the other being free-trade. It is hypothesized that each economy can be compared under autarky independently of the other. For profitable trade between such hypothetical economies to take place, there must be some difference between the autarky or pre-trade prices. The various theories of comparative advantage proceed to postulate a variety of determinants of the autarky price differentials. There are at least three ways in which the autarky construct in the theory of comparative advantage is problematic.

First, when capital is treated as produced means of production in the economic model, either the conventional interpretation of the factor proportions theory of comparative advantage requires modification, or the autarky construct itself must be altered. The modern neo-Ricardian theory of comparative advantage takes the former route, arguing that when capital as produced means of production is included, independent determinants of income distribution must be added to the Ricardian theory of comparative advantage, which is based on technological differences between economies. (For an introduction to neo-Ricardian trade theory, see Steedman (ed.), 1979; see also Metcalfe and Steedman 1981.) Neo-classical and some neo-Marxian interpretations of the theory of comparative advantage argue that the autarky construct itself must be modified to allow autarky comparisons at common prices between non-trading economies for the consistent valuation of produced means of production. The Heckscher–Ohlin– Samuelson factor proportions theory can then be applied in situations where there are produced means of production. The latter resolution of the difficulties created by produced means of production in the theory of comparative advantage is consistent with the empirical observation that no real economy is ever observed without trade or independently of other economies. In practical terms, the modified concept autarky really means ‘less trade’ rather than ‘no-trade’. (See Ethier 1981, and Smith 1984, for a discussion of the neo-classical response to these issues.)

Second, the measurement of autarky price differentials when there are more than two commodities is not unambiguous. For a recent discussion of this problem and a generalization of the principle of comparative advantage, see Deardorff (1980).

A third problem with the use of the autarky construct in the theory of comparative advantage is that it is often defined in the context of an economic model which has no descriptive content. Whilst this is perfectly legitimate for some

theoretical work, for empirical and policy purposes it is not appropriate to leave out a description of either the level of development of the forces of production, the relations of production which pertain to the economy or economics under consideration, or the superstructural arrangements in place. (For a modern re-statement of Marx's theory of history, see Cohen 1978, 1983; see also Dobb 1973.) This can lead to an over-emphasis on the role of the market and static efficiency criteria, and an understatement of dynamic and institutional factors, in the resultant theories of comparative advantage. (For a discussion of some of these issues, see Evans and Alizadeh 1984.) The latter observation has an important bearing on the importance of autarky or self-sufficiency as a trading policy.

The main 19th-century exponents of economic self-sufficiency were Hamilton and List. They did not advocate autarky in the literal sense used in the theory of comparative advantage, but they did argue that new industrialized nations required protection of their infant industries before free trade could be embarked upon. There have been many 20th-century counterparts of List and Hamilton. In the 1920s, Preobrazhensky argued for import-substituting industrialization financed by the taxation of agriculture in a process called primitive socialist accumulation (for a formal statement of Preobrazhensky's problem, see Bardhan 1970, ch. 9). In the early postwar period. Prebisch and Singer had a powerful influence on the Economic Commission for Latin America (ECLA). arguing for import-substituting industrialization to offset hypothesized adverse terms of trade movements and adverse monopoly conditions facing primary commodity producers. (By now, there is strong statistical evidence to support the Prebisch–Singer hypothesis on the declining trend of the net barter terms of trade between primary commodities and manufactures (excluding oil) for the whole of the 20th century; see Sapsford 1985.) In Eastern Europe, in spite of considerable integration of their national economies, drives towards import substitution and self-sufficiency both nationally and as a trading bloc remain powerful tendencies in their economic mechanisms. Until recently, China followed a policy of near autarky, and only in the 1970s have there been moves in India to begin to dismantle powerful barriers to trade. An offshoot of the ECLA school, with strong neo-Marxian influences, has argued that by remaining open to the world economy, developing economies will not develop but will suffer a process of underdevelopment through mechanisms of dependency and unequal exchange. Amin, the leading advocate of a semi-autarkic development strategy, bases his argument on a model of unequal exchange which is not well founded theoretically or empirically. (For a statement and critique of Amin's theory of unequal exchange, see Amin 1973, and Evans 1981 and 1984.)

Some of the theoretical arguments for interfering with market mechanism, with direct and indirect consequences for the pattern and extent of trade, are agreed by all schools of thought. The presence of externalities and strong economies of learning combined with varying degrees of market distortion and market failure provide the basis for the modern theory of domestic market distortions. Within this context, protection through intervention in trade is likely to be worse than subsidies or other policies aimed directly towards policy objectives.

Increasingly, neoclassical economists and some of the main international agencies such as the World Bank and the International Monetary Fund, with a sharply enhanced policy role through the conditionality attached to debt re-negotiation agreements, argue that the development of dynamic comparative advantage is better served by imperfect markets than imperfect governments. This perspective is strongly disputed by many who stress the importance of state and parastatal institutions operating in conjunction with the market mechanism, often in the context of a rapidly growing national capitalist class and national capitalist firms (for an overview of some of these arguments, see Kaplinsky 1984). Different views on the length of the learning period and the length of time for which it is appropriate for the state to be the driving force in a national development strategy lie behind the important policy debates on the role of freer trade and the world market in all economies, east, west and south. In the latter part of the 1980s, many

debt-ridden developing countries are being asked to trade their way out of debt in the context of a sluggish and closing world economy, often with disastrous domestic consequences for the poorest and weakest citizens in their midst. The few countries which have a choice in the matter must find their sources of growth in their internal markets rather than through trade.

It is not easy to assess the degree of success of policies of economic self-sufficiency. Whilst a strong case can be made for greater economic self-sufficiency as a part of the process of developing a national economy, it is not clear how long such a policy should be carried on, or how selective government policies should be towards the protection of different industries. In practice, the remarkable growth performance of many developing countries in the postwar period has been achieved in very widely differing circumstances and with greater or less economic self-sufficiency. What is clear is that, excepting the special cases of some small city states, all late developers have gone through periods of development of their national economies with policies of greater economic self-sufficiency. Only in the very extreme cases of economic self-sufficiency, such as pursued in Cambodia under the Khmer Rough in the 1970s, or of extreme protection, such as in Ghana, can it be said unequivocally that the drive for economic self-sufficiency, foregoing the static gains from trade, has contributed decisively to subsequent economic disaster.

See Also

- ▶ Comparative Advantage
- ▶ Free Trade and Protection
- ▶ Heckscher–Ohlin trade Theory
- ▶ National System

Bibliography

Amin, S. 1973. *L'échange inégal et la loi de la valeur: la fin d'un débat*. Paris: Editions Anthropus – IDFP.

Bardhan, P.K. 1970. *Economic growth, development and foreign trade: A study in pure theory*. New York: Wiley.

Cohen, G.A. 1978. *Karl Marx's theory of history: A defence*. Oxford: Oxford University Press.

Cohen, G.A. 1983. Forces and relations of production. In *Marx: A hundred years on*, ed. B. Mathews. London: Lawrence & Wishart.

Deardorff, A.V. 1980. The general validity of the law of comparative advantage. *Journal of Political Economy* 88: 941–957.

Dobb, M. 1973. *Theories of value and distribution since Adam Smith*. Cambridge: Cambridge University Press.

Ethier, W. 1981. A reply to Professors Metcalfe and Steedman. *Journal of International Economies* 11: 273–277.

Evans, H.D. 1981. Trade, production and self-reliance. In *Dependency theory: A critical assessment*, ed. D. Seers. London: Frances Pinter.

Evans, H.D. 1984. A critical assessment of some neo-marxian trade theories. *Journal of Development Studies* 20(2): 202–226.

Evans, H.D., and P. Alizadeh. 1984. Trade, industrialisation and the visible hand. *Kaplinsky* 1984.

Kaplinsky, R. (ed.) 1984. *Third World industrialisation in the 1980's: Open economics in a closing World*. Special Issue of *The Journal of Development Studies*, October.

Metcalfe, J.S., and I. Steedman. 1981. On transformation of theories. *Journal of International Economics* 11: 267–271.

Sapsford, D. 1985. The statistical debate on the net barter terms of trade between primary commodities and manufactures; a comment and some additional evidence. *Economic Journal* 95: 781–788.

Smith, A. 1984. Capital theory and trade theory. In *Handbook of international economics*, vol. 1, ed. R.W. Jones and P.B. Kenen, 289–324. Amsterdam: North-Holland.

Steedman, I. (ed.). 1979. *Fundamental issues in trade theory*. London: Macmillan.

Autonomous Expenditures

M. Sawyer

Abstract

The idea of autonomous expenditures is usually associated with a simple Keynesian model of the economy and refers to those expenditures which are treated as exogenously given within the context of the model being used. The contrast is drawn between autonomous expenditures and induced expenditures. Autonomous expenditures are those which are unrelated to the other economic variables

being considered, though it is income which is generally taken to be the key economic variable which does not influence autonomous expenditures. Induced expenditures are influenced by other economic variables, with the level of income being a major influence.

The idea of autonomous expenditures is usually associated with a simple Keynesian model of the economy and refers to those expenditures which are treated as exogenously given within the context of the model being used. The contrast is drawn between autonomous expenditures and induced expenditures. Autonomous expenditures are those which are unrelated to the other economic variables being considered, though it is income which is generally taken to be the key economic variable which does not influence autonomous expenditures. Induced expenditures are influenced by other economic variables, with the level of income being a major influence.

In the simplest formation of a Keynesian model, consumption expenditure is taken as $a + c.\ Y$ where Y is the level of income, and c the marginal propensity to consume out of income with a value of less than unity, and investment expenditure is taken as I (fixed). Then total expenditure equals $a + I + c.\ Y$. The component $a + I$ is the autonomous expenditure and $c.Y$ induced expenditure. In equilibrium, with income equal to expenditure, then $Y = a + I + c.\ Y$, so that $Y = (a + I)/(1 - c)$. This formula indicates the potential importance of autonomous expenditure in that it is autonomous expenditure which determines the level of income. Changes in autonomous expenditure are predicted to lead to changes in income.

Outside the simple model outlined above, the allocation of expenditure into categories of 'autonomous' and 'induced' is not straightforward. The difficulties which arise can be examined under two main headings. First, there will generally be some lags between the receipt of income and its effects on expenditure. A rise in income in the current period may have effects on expenditure in a number of future periods. Within the current period, there will be some expenditure induced by current income, some by previous income and some will be autonomous. Second, there are many categories of expenditure, besides consumption expenditure and investment, and these expenditures may be difficult to categorize as between induced and autonomous expenditures. Investment itself (and particularly investment in stocks and work-in-progress) may be related to income and hence partially induced expenditure. Government expenditure can vary automatically though inversely with the level of income (e.g. unemployment benefits), whilst taxation generally rises with income. But elements of government expenditure and taxation may be varied by the government in response to the level of income (particularly if the government was operating a Keynesian demand-management policy).

The simple Keynesian approach to macroeconomics paid particular attention to the importance of autonomous expenditures in the determination of the level of income. But in order to test that approach, different types of expenditure have to be classified as autonomous or induced. An attempt to do this, and to contrast the Keynesian approach with a monetarist approach, was made by Friedman and Meiselman (1963). The conclusion of that article was challenged by Ando and Modigliani (1965) and by de Prano and Mayer (1965). These articles did not reach any shared conclusions, but they did indicate the difficulties of making the concept of autonomous and induced expenditures operational (as well as raising a number of other methodological issues).

The actual definition of autonomous expenditure used by Friedman and Meiselman (1963) was the sum of net private domestic investment, government deficit on current account and the net foreign balance. These authors arrive at their definitions of autonomous expenditure by reference to the statistical relationship between various possible definitions of autonomous expenditure. Ando and Modigliani (1965) define autonomous expenditure as investment, exports, most government expenditure minus property taxes. Thus they regard most taxes as induced rather than autonomous. But they arrive at their definitions by their subjective views rather than the formal statistical tests applied by Friedman and Meiselman (which they reject).

The importance of the distinction drawn between autonomous and induced expenditures is threefold. First, it established a break with Say's Law that supply creates its own demand. In effect, under Say's Law, all expenditure is induced, so that an increase in the supply of goods and services would generate income for the suppliers, which in turn leads to a rise in demand. There could be some temporary disruption from such changes, but no prolonged effect from a discrepancy between supply and demand since the income generated for the suppliers is all spent. When the distinction between autonomous and induced expenditures is made, it is implicit that induced expenditures are less than total income and that the difference between the level of income which would be generated at full employment and the corresponding induced expenditures would not necessarily be filled by autonomous expenditures. Thus demanddeficient unemployment would then arise.

Second, there is usually an approximate identification of autonomous expenditures with investment, exports and government expenditures. It is a relatively short step to consider the different types of autonomous expenditures as substitutes for one another in the sense of contributing the same effect to the level of aggregate demand. In policy terms, this clearly would lead to suggestions that variations in government expenditure be used to offset fluctuations in private autonomous expenditure, particularly investment, in order to limit the extent of business fluctuations.

Third, autonomous expenditure is seen as the active ingredient in the level of aggregate demand, whilst induced expenditure is viewed as passive. Thus, induced expenditures are seen as adjusting passively to the level of income (as indicated above), whereas variations in autonomous expenditure are seen as leading to variations in the level of income (and in induced expenditure).

Bibliography

Ando, A., and F. Modigliani. 1965. The relative stability of monetary velocity and the investment multiplier. *American Economic Review* 55(4): 693–728.

De Prano, M., and T. Mayer. 1965. Tests of the relative importance of autonomous expenditure and money. *American Economic Review* 55(4): 729–752.

Friedman, M. and Meiselman, D. 1963. The relative stability of monetary velocity and the investment multiplier in the United States, 1897–1958. In *Commission on money and credit*, *stabilization policies*, ed. E. Carey Brown et al., 165–268. Englewood Cliffs: Prentice-Hall.

Autoregressive and Moving-Average Time-Series Processes

Marc Nerlove

Characterization of time series by means of autoregressive (AR) or moving-average (MA) processes or combined autoregressive moving-average (ARMA) processes was suggested, more or less simultaneously, by the Russian statistician and economist, E. Slutsky (1927), and the British statistician G.U. Yule (1921, 1926, 1927). Slutsky and Yule observed that if we begin with a series of purely random numbers and then take sums or differences, weighted or unweighted, of such numbers, the new series so produced has many of the apparent cyclic properties that are thought to characterize economic and other time series. Such sums or differences of purely random numbers are the basis for ARMA models of the processes by which many kinds of economic time series are assumed to be generated, and thus form the basis for recent suggestions for analysis, forecasting and control (e.g., Box and Jenkins 1970).

Let L be the lag operator such that $L^k x_t = x_{t-k}$. Consider the familiar pth order linear, homogeneous, deterministic difference equation with constant coefficients common in discrete dynamic economic analysis (e.g. Chow 1975)

$$\psi(L)y_t = 0$$

Or

$$y_i - \psi_1 y_{t-1} - \cdots - \psi_p y_{t-p} = 0. \qquad (1)$$

Relationships are seldom exact, however, so we introduce a serially uncorrelated random shock ε_t with zero mean and constant variance:

$$E \in_t = 0$$
$$E \in_t \in_{t'} = \begin{cases} \sigma^2, & t = t' \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Thus

$$\psi(L) y_t = \varepsilon_t, \quad (3)$$

which is the pth-order autogressive process, AR(p), with constant coefficients studied by Yule (1927).

If the stochastic term in (3) is itself assumed to be a linear combination of past values of a variable such as ε_t with properties (2), for example,

$$\psi(L) y_t = \mu_t, \quad (4)$$

where $\mu_t = \phi(L)\varepsilon_t$, and $\phi(\cdot)$ is a polynomial of order q, then the process is a mixed autoregressive moving-average process of order (p, q), ARMA (p, q). The process generating μ_t is simply a moving-average process of order q, MA(q).

Such dynamic processes, under appropriate conditions on the coefficients of ψ, ϕ, and the distribution of ε_t have found wide application in both theoretical and empirical economics. The ability of such processes to describe the evolution of a series has made ARMA models a powerful tool for forecasting economic time series and other applications, such as seasonal adjustment. Moreover, because the models can capture a wide range of stochastic properties of economic time series, they have been widely used in models involving rational expectations (e.g. Whiteman 1983).

Univariate Arma Models

Conditions for weak stationarity and invertibility (i.e. capability of being expressed as a pure, but possibly infinite, AR) are most easily discussed in terms of the so-called z-transform or autocovariance-generating transform of the model. This is obtained for models (3) and (4) by replacing the lag operator by a complex variable z; thus, in general,

$$B(z) = \frac{\phi(z)}{\psi(z)}, \quad (5)$$

where the expression on the right converges. If the roots of $\psi(z) = 0$ do not lie strictly outside the unit circle (i.e. some lie on or inside), the process described by (3) or (4) will not be stationary, nor will the expression on the right converge outside of a circle with radius less than one. (*See* ▶ Time Series Analysis.) In order to find a purely AR representation of MA and ARMA models, we require that $1/\phi(z)$ converge in the same region, so that $\phi(z) = 0$ must also have roots outside the unit circle. In this case, $B(z)$ is well-defined everywhere outside the unit circle, and the model defined by (4) is both weakly stationary and invertible; the representation

$$y_t = B(L)\varepsilon_t \quad (6)$$

is a one-sided, infinite-order MA, with $\sum_{j=-\infty}^{\infty} b_j^2 < \infty$.

In Wold (1938) it is shown that every discrete weakly stationary process may be decomposed into a purely linearly deterministic part (which can be predicted exactly from a sufficient past history) and a part which corresponds to (6) above. (See the discussions of stationarity and ergodicity in ▶ Time Series Analysis.)

Let the autocovariances of a stationary, zero-mean time series, x_t, be given by

$$\gamma(\tau) = E x_t x_{t-\tau}. \quad (7)$$

The function

$$g(z) = \sum_{\tau=-\infty}^{\infty} z^r \gamma(\tau) \quad (8)$$

is called the autocovariance generating function. If the function $g(z)$ is known and analytic in a certain region, it is possible to read off the autocovariances of the time series as the coefficients in a Laurent series expansion of the function there.

For a linearly nondeterministic time series with one-sided MA representation (6) the autocovariance generating transform is given by

$$g_{yy}(z) = \sigma^2 B(z) B\left(z^{-1}\right). \tag{9}$$

This function is analytic everywhere in an annulus about the unit circle. If y_t is generated by a stationary ARMA model with invertible MA component then $g_{yy}(z)$ will have no zeros anywhere in this annulus. On the unit circle itself the *spectral density* of the series is proportional by a factor of $(2\pi)^{-1}$ to the autocovariance generating transform:

$$f_{yy}(\lambda) = (1/2\pi) g_{yy}\left(e^{i\lambda}\right), \quad -\pi \le \lambda < \pi. \tag{10}$$

Stationary, invertible ARMA processes give rise to time series with spectral densities which are strictly positive in the interval $(-\pi, \pi)$.

Let $1/\beta_j, j = 1,\ldots, q$ be the roots, not necessarily distinct, of $\phi(z) = 0$, and $1/\alpha_j, j = 1,\ldots, p$ be the roots of $\psi(z) = 0$. For a stationary, invertible ARMA model all these roots lie outside the unit circle. The autocovariances of the time series generated by this model are

$$\begin{aligned}\gamma(\tau) &= (1/2\pi i)\oint_{|z|=1} z^{-\tau-1} g(z)\mathrm{d}z \\ &= (\sigma^2/2\pi i)\oint_{|z|=1} z^{p+|\tau|-q-1} \\ &\times \left\{\prod_{j=1}^{q}\left(1-\beta_j z\right)\left(z-\beta_j\right)/\prod_{k=1}^{p}(1-\alpha_k z)(z-\alpha_k)\right\} \\ &\mathrm{d}z.\end{aligned} \tag{11}$$

By the residue theorem, the integral on the right-hand side of (11) is $2\pi i$ times the sum of the residues enclosed by the unit circle. This fact allows a particularly simple calculation of the autocovariances of a time series generated by an ARMA model (see Nerlove et al. 1979, pp. 78–85). For example, for the general pth-order autoregression, AR(p), with distinct roots, the result is

$$\begin{aligned}&\gamma(\tau) \\ &= \sum_{k=1}^{p}\left[\alpha_k^{p+|\tau|-1} \Big/ \left\{\prod_{j=1}^{p}(1-\alpha_j a_k)\prod_{\substack{j=1\\ j\neq k}}^{p}(\alpha_j-\alpha_k)\right\}\right],\end{aligned} \tag{12}$$

and for the ARMA(1, 1) model, it is

$$\begin{aligned}\gamma(\tau) &= \alpha^{|\tau|}(1-\alpha\beta)(1-\beta/\alpha)/(1-\alpha^2), \\ &\qquad \tau = \pm 1, \pm 2, \ldots \\ &= \left(1+\beta^2-2\alpha\beta\right)/(1-\alpha^2), \\ &\qquad \tau = 0.\end{aligned} \tag{13}$$

Formulation and Estimation of Univariate Arma Models

The problem of *formulating* an ARMA model refers to determination of the orders p and q of the AR and MA components, while the *estimation* problem is that of determining the values of the parameters of the model, for example, the roots $1/\alpha_j, j = 1,\ldots, p$, and $1/\beta_k, k = 1,\ldots, q$, and the variance σ^2 of ε_t.

Box and Jenkins (1970), among others, have suggested the use of the sample autocorrelation and partial autocorrelation functions as an approach to the problem of formulating an ARMA model. It is known, however, that the estimates of these functions are poorly behaved relative to their theoretical counterparts and, thus, provide a somewhat dubious basis for model formulation (Nerlove et al. 1979, pp. 57–68, 105–106; Hannan 1960, p. 41).

More recently, information-theoretic approaches to model formulation, having a rigorous foundation in statistical information theory, have been proposed. These procedures are designated for order determination in general ARMA (p, q) models. The Akaike (1973) Information Criterion (AIC) leads to selection of the model for which the expression:

$$\mathrm{AIC}(k) = \ln \hat{\sigma}^2_{\mathrm{ML}} + 2k/T \tag{14}$$

is minimized, where $\hat{\sigma}^2_{\mathrm{ML}}$ is the maximum likelihood estimate of σ^2_ε, T is sample size, and $k = p + q$. It is well known that the AIC is not consistent, in the sense that it does not lead to selection of the correct model with probability one in large samples (Shibata 1976; Hannan and Quinn 1979; Hannan 1980; Kashyap 1980). The procedure does, however, have special benefits when selecting the order of an AR model, as

shown by Shibata (1980). Specifically, he shows that if the true model can *not* be written as a finite AR, but an AR is fitted anyway, then use of the AIC minimizes asymptotic mean-squared prediction error within the class of AR models.

Schwarz (1978) and Rissanen (1978) develop a consistent modification of the AIC which has become known as the Schwarz Information Criterion (SIC). This criterion selects the model which minimizes:

$$\mathrm{SIC}(k) = \ln \hat{\sigma}^2_{\mathrm{ML}} + \frac{\ln T}{T}(k), \qquad (15)$$

and Hannan (1980) shows that this procedure identifies the true model with probability one in large samples, so long as the maximum possible orders of the AR and MA components are known.

Once the orders p and q are determined, the problem of *estimating* the parameters of the ARMA model remains. Various approaches in the time domain are available, such as least squares, approximate maximum likelihood (Box and Jenkins 1970), or exact maximum likelihood (Newbold 1974; Harvey and Philips, 1979; Harvey 1981). Approximate maximum likelihood in the frequency domain is also possible (Hannan 1969b; Hannan and Nicholls 1972; Nerlove et al. 1979, pp. 132–6). The latter is based upon the asymptotic distribution of the sample periodogram ordinates.

Estimation of pure AR models (no MA component) is particularly simple since ordinary least squares yield consistent parameter estimates. The basis of such estimation is the set of Yule–Walker equations (Yule 1927; Walker 1931). Consider the AR(p) process;

$$y_t = \sum_{i=1}^{p} \psi_i y_{t-i} + \varepsilon_t. \qquad (16)$$

Multiplying (16) by $y_{t-\tau}$, $\tau \geq 0$, taking expectations, and recognizing that $\gamma(\tau) = \gamma(-\tau)$ gives

$$\gamma(\tau) = \sum_{i=1}^{p} \psi_i \gamma(\tau - i), \quad \tau > 0. \qquad (17)$$

Dividing (17) by the variance $\gamma(0)$, we obtain the system of Yule–Walker equations:

$$\rho(\tau) = \sum_{i=1}^{p} \psi_i \rho(\tau - i), \quad \tau > 0, \qquad (18)$$

which relate the autocorrelations of the process. This pth-order linear system is easily solved for the ψ_i, $i = 1,\ldots, p$, in terms of the first p autocorrelations. In practice, the theoretical autocorrelations are replaced by their sample counterparts, yielding estimates of the ψ_i, $i = 1,\ldots, p$. These parameter estimates may be conveniently used as start-up values for the more sophisticated, iterative estimation procedures discussed above.

Estimation of MA or mixed models by exact maximum likelihood methods is complicated further by a tendency to obtain a local maximum of the likelihood function at a unit root of the MA component, even when no roots are close to the unit circle (Sargan and Bhargava 1983; Anderson and Takemura 1984).

Prediction

Optimal linear least squares prediction of time series generated by ARMA processes may be obtained for known parameter values by the Wiener–Kolmogorov approach (Whittle 1983). If y_t is generated by a stationary, invertible ARMA model with one-sided MA representation (6), a very simple expression may be given for the linear minimum meansquare error (MMSE) prediction of y_{t+v} at time t, y^*_{t+v}, in terms of its own (infinite) past

$$y^*_{t+v} = C(z) y_t, \qquad (19)$$

where

$$C(z) = \sum_{j=0}^{\infty} c_j z^j = \frac{1}{B(z)} \left[\frac{B(z)}{z^v} \right]_+ .$$

The operator $[.]_+$ eliminates negative powers of z.

Suppose that y_t is AR(1):

$$y_t = \alpha y_{t-1} + \varepsilon_t, \quad |\alpha| < 1,$$

then $y^*_{t+v} = \alpha^v y_i$. If y_t is AR(2):

$$y_t = (\alpha_1 + \alpha_2) y_{t-1} - \alpha_1 \alpha_2 y_{t-2} + \varepsilon_t, |\alpha_1|, |\alpha_2| < 1,$$

then $y^*_{t+1} = (\alpha_1 + \alpha_2) y_t - \alpha_1 \alpha_2 y_{t-1}$. In general, the result for AR(p) as in (1) is $y^*_{t+v} = \psi_1 y^*_{t+v-1} + \cdots + \psi_p y^*_{t+v-p}$, where $y^*_{t-j} = y_{t-j}$, for $j = 0, 1, \ldots$, at time t. Thus for pure autoregression the MMSE prediction is a linear combination of only the p most recently observed values.

Suppose that y_t is MA(1):

$$y_t = \varepsilon_t - \beta \varepsilon_{t-1}, \quad |\beta| < 1,$$

then $y^*_{t+1} = -\beta \sum_{j=0}^{\infty} \beta^j x_{t-j}$ and $y^*_{t+v} = 0$ for all $v > 1$. For moving-average processes, in general, predictions for a future period greater than the order of the process are zero and those for a period less distant cannot be expressed in terms of a finite number of past observed values.

Finally, suppose that y_t is ARMA(1, 1): $y_t - \alpha y_{t-1} = \varepsilon_t - \beta \varepsilon_{t-1}, |\alpha|, |\beta| < 1$, then $y^*_{t+v} = \alpha^{v-1} (\alpha - \beta) \sum_{j=0}^{\infty} \beta^j y_{t-j}$. For further examples, see Nerlove et al. 1979, pp. 89–102.

When an infinite past is not available and the parameter values of the process are not known, the problem of optimal prediction is more complicated. The most straightforward approach is via the state-space representation of the process and the Kalman filter (Kalman 1960; Meinhold and Singpurwalla 1983).

Multivariate Arma Processes

Let $\psi(\cdot)$ and $\Phi(\cdot)$ be $K \times K$ matrix polynomials in the lag operator, y_t and ε_t be $K \times 1$ vectors. Then the K-variate ARMA(p, q) process is defined as

$$\psi(L) y_t = \Phi(L) \varepsilon_t, \quad \varepsilon_t^{iid}(0, \sum), \tag{20}$$

where $\Psi(L) = \Psi_0 - \Psi_1 L - \ldots - \Psi_p L^p$ and $\Phi(L) = \Phi_0 - \Phi_1 L - \ldots - \Phi_q L^q$, with each ψ_j and Φ_j, $y = 0, 1, \ldots$, being a $K \times K$ matrix. The model is weakly stationary if all the zeros of det $|\Psi(z)|$ lie outside the unit circle (Hannan 1970), and invertible if all the zeros of det $|\Phi(z)|$ also do.

In addition to the issues of formulation, estimation, and prediction, which arise in the univariate case as well, identification (in the usual econometric sense) becomes an important problem. Hannan (1969a) shows that a stationary vector AR process is identified if Φ_0 is an identity matrix (i.e. no instantaneous coupling), and Ψ_p is nonsingular. Hannan (1971) extends the analysis to recursive systems and systems with prescribed zero restrictions.

There are three approaches to the formulation of multivariate ARMA models. Nerlove et al. (1979) and Granger and Newbold (1977) develop an augmented single-equation procedure, and Wallis (1977) and Wallis and Chan (1978) develop another procedure which involves preliminary univariate analysis.

A second approach is due to Tiao and Box (1981), who use multivariate analogues of the autocorrelation and partial autocorrelation functions as a guide to model formulation. Their approach is computationally quite simple and usually leads to models with a tractable number of parameters. Identification is achieved by allowing no instantaneous coupling among variables.

Finally, the information-theoretic model formulation procedures which were discussed above generalize to the multivariate ARMA case. Quinn (1980) shows that Schwarz's criterion (SIC) again provides a consistent estimate of the vector AR order. In a large Monte Carlo comparison of criteria for estimating the order of a vector AR process (VAR), Lütkepohl (1985) shows the clear superiority of the SIC in medium-sized samples; the SIC chooses the correct model most often and leads to the best forecasting performance.

As in the case of univariate ARMA models, estimation in the multivariate case may be carried out in the frequency domain (Wilson 1973; Dunsmuir and Hannan 1976) or in the time domain (Hillmer and Tiao 1979). An exact likelihood function in the time domain may also be derived by the Kalman filter by casting the multivariate ARMA model in state space form. Anderson (1980) provides a good survey of estimation in both time and frequency domains.

Prediction in the multivariate case with an infinite past is a straightforward generalization of the results for the univariate case (Judge et al. 1985, pp. 659–60). When only a finite past is available and the parameters of the process must be estimated, the most straightforward approach is again through the Kalman filter (see also Yamamoto 1981).

Applications

In addition to their obvious uses in forecasting, ARMA models, especially multivariate ARMA models, have a wide range of economic and econometric application.

The use of time-series methods in formulating distributed-lag models is discussed at length in Nerlove (1972) and Nerlove et al. (1979, pp. 291–353) and applied in the latter to an analysis of US cattle production. The notion of quasi-rational expectations introduced there is that the expectations on the basis of which economic agents react may, under certain conditions, be assumed to be the statistical expectations of the variables in question, conditional on observations of past history. If these variables are generated by time-series processes, such as those discussed in this entry, time-series methods may be used to derive expressions for the MMSE forecasts for any relevant future period; these MMSE forecasts are, by a well-known result, the aforementioned conditional expectations.

An econometric definition of causality based on time-series concepts has been developed by Granger (1969) and extended by Sims (1972). Let (x_t, y_t) be a pair of vectors of observations on some economic time series, and let Ω_{t-1} be the information available up to time t, which includes $\{(x_{t-1}, y_{t-1}), (x_{t-2}, y_{t-2}), \ldots\}$. Granger gives the following definitions in terms of the conditional variances:

Definition 1 x causes y if and only if

$$\sigma^2(y_t|\Omega_{t-1}) < \sigma^2(y_t|\Omega_{t-1} - \{x_{t-1}, x_{t-2}, \ldots\}),$$

where $\Omega_{t-1} - \{x_{t-1}, x_{t-2}, \ldots\}$ is the information set omitting the past of the series x_t.

Definition 2 x causes y instantaneously if and only if

$$\sigma^2(y_t|\Omega_{t-1}, x_t) < \sigma^2(y_t|\Omega_{t-1}).$$

It may happen that both x causes y, and y causes x; then x; and y are related by a feedback system. In applications (x_t, y_t), is generally assumed to be generated by multivariate ARMA processes, and Ω_t is assumed to consist only of the past history of (x_t, y_t). Since ARMA models are applicable only to weakly stationary time series it must further be assumed that any transformation necessary to achieve stationarity is causality preserving. Granger's (1969) test for causal association is based on a multivariate AR representation, while Sims (1972) bases his on an equivalent MA representation. Sims also introduces a regression-based test related to the above which makes use of both future and past values of the series x_t in relation to the current value of y_t. Pierce and Haugh (1977) show that causality may also be tested in univariate representations of the series. Feige and Pierce (1979) and Lütkepohl (1982) show that the direction of causality so defined may be sensitive to the transformations used to achieve stationarity, and to the definition of the information set.

Time series methods have also been applied to the analysis of the efficiency of capital markets (Fama 1970). The question is whether market prices fully reflect available information, for example, in a securities market. Efficiency requires that the relevant information set be that actually used by the market participants. Since the latter is inherently unobservable, tests of the efficiency of a market can be carried out only within the context of a particular theory of market equilibrium. Various alternatives lead to tests, based on AR or more general models, of the rates of return for different securities over time in the presence of shocks of various sorts which may or may not represent the introduction of new information (Ball and Brown 1968; Fama et al. 1969; Scholes 1972).

Finally, an important example of the use of time-series methods in econometrics has been put forth in the controversial revisionist views of Sargent and Sims (1977), and Sims (1980) on

appropriate methods of econometric modelling. These views may be traced back to the work of T. C. Liu (1960) who argued that when only reliable *a priori* restrictions were imposed, most econometric models would turn out to be underidentified; furthermore, he argued that most of the exclusion restrictions generally employed, and the assumptions about serial correlation made to justify treating certain lagged values of endogenous variables as predetermined, were invalid; he concluded that only unrestricted reduced form estimation could be justified. The revisionist approach treats *all* variables as endogenous and, in general, places no restrictions on the parameters except the choice of variables to be included and lengths of lags. Attention in this approach is focused on the estimation of a general relationship among a relatively short list of variables rather than policy analysis and structural inference, which have been the emphasis of mainstream econometrics. As such, the approach has been mainly useful for data description and forecasting.

See Also

- ARIMA Models
- Econometrics
- Ergodic Theory
- Estimation
- Multivariate Time Series Models
- Spectral Analysis
- Stationary Time Series
- Time Series Analysis

Bibliography

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, ed. B.N. Petrov and F. Csaki, 267–287. Budapest: Akademiai Kiado.

Anderson, T.W. 1980. Maximum likelihood estimation for vector autoregressive moving average models. In *Directions in time series*, ed. D.R. Brillinger and G.C. Tiao, 49–59. Hayward: Institute of Mathematical Statistics.

Anderson, T.W., and Takemura, A. 1984. *Why do noninvertible moving averages occur?* Technical report no. 13, Department of Statistics, Stanford University.

Ball, R., and P. Brown. 1968. An empirical evaluation of accounting income numbers. *Journal of Accounting Research* 6: 159–178.

Box, G.E.P., and G.M. Jenkins. 1970. *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Chow, G.C. 1975. *Analysis and control of dynamic economic systems*. New York: Wiley.

Dunsmuir, W.T.M., and E.J. Hannan. 1976. Vector linear time series models. *Advances in Applied Probability* 8(2): 339–364.

Fama, E.F. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25(2): 383–471.

Fama, E.F., M. Jensen, L. Fisher, and R. Roll. 1969. The adjustment of stock market prices to new information. *International Economic Review* 10(1): 1–21.

Feige, E.L., and D.K. Pierce. 1979. The casual causal relation between money and income: Some caveats for time series analysis. *The Review of Economics and Statistics* 61(4): 521–533.

Granger, C.W.J. 1969. Investigating causal relationships by econometric models and cross-spectral methods. *Econometrica* 37(3): 424–438.

Granger, C.W.J., and P. Newbold. 1977. *Forecasting economic time series*. New York: Academic.

Hannan, E.J. 1969a. The identification of vector mixed autoregressive-moving average systems. *Biometrika* 56(1): 223–225.

Hannan, E.J. 1969b. The estimation of mixed moving average autoregressive systems. *Biometrika* 56(3): 579–593.

Hannan, E.J. 1970. *Multiple time series*. New York: Wiley.

Hannan, E.J. 1971. The identification problem for multiple equation systems with moving average errors. *Econometrica* 39(5): 751–765.

Hannan, E.J. 1980. The estimation of the order of an ARMA process. *Annals of Statistics* 8(5): 1071–1081.

Hannan, E.J., and D.F. Nicholls. 1972. The estimation of mixed regression, autoregression, moving average and distributed lag models. *Econometrica* 40(3): 529–547.

Hannan, E.J., and B.G. Quinn. 1979. The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B* 41(2): 190–195.

Harvey, A.C. 1981. *Time series models*. Oxford: Philip Allan.

Harvey, A.C., and G.D.A. Phillips. 1979. The estimation of regression models with ARMA disturbances. *Biometrika* 66(1): 49–58.

Hillmer, S.C., and G.C. Tiao. 1979. Likelihood function of stationary multiple autoregressive moving average models. *Journal of the American Statistical Association* 74(367): 652–660.

Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl, and T.C. Lee. 1985. *The theory and practice of econometrics*, 2nd ed. New York: Wiley.

Kalman, R.E. 1960. A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering* 82D: 35–45.

Kashyap, R.L. 1980. Inconsistency of the AIC rule for estimating the order of AR models. *IEEE Transactions on Automatic Control* 25(5): 996–998.
Liu, T.C. 1960. Underidentification, structural estimation, and forecasting. *Econometrica* 28(4): 855–865.
Lütkepohl, H. 1982. Non-causality due to omitted variables. *Journal of Econometrics* 19: 367–378.
Lütkepohl, H. 1985. Comparison of criteria for estimating the order of a vector autoregressive process. *Journal of Time Series Analysis* 6(1): 35–52.
Meinhold, R.J., and N.D. Singpurwalla. 1983. Understanding the Kalman filter. *American Statistician* 37: 123–127.
Nerlove, M. 1972. Lags in economic behaviour. *Econometrica* 40(2): 221–251.
Nerlove, M., D.M. Grether, and J.L. Carvalho. 1979. *Analysis of economic time series: A synthesis*. New York: Academic.
Newbold, P. 1974. The exact likelihood function for a mixed autoregressive-moving average process. *Biometrika* 61(3): 423–426.
Pierce, D.A., and L.D. Haugh. 1977. Causality in temporal systems: Characterizations and a survey. *Journal of Econometrics* 5(3): 265–293.
Quinn, B.G. 1980. Order determination for a multivariate autoregression. *Journal of the Royal Statistical Society, Series B* 42(2): 182–185.
Rissanen, H. 1978. Modelling by shortest data description. *Automatica* 14(5): 465–471.
Sargan, J.D., and A. Bhargava. 1983. Maximum likelihood estimation of regression models with moving average errors when the root lies on the unit circle. *Econometrica* 51(3): 799–820.
Sargent, T.J., and C.A. Sims. 1977. Business cycle modeling without pretending to have too much a priori economic theory. In *New methods of business cycle research*, ed. C.A. Sims. Minneapolis: Federal Reserve Bank of Minneapolis.
Scholes, M. 1972. The market for securities: Substitution versus price pressure and the effects of information on share prices. *Journal of Business* 45(2): 179–211.
Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6(2): 461–464.
Shibata, R. 1976. Selection of the order of an autoregressive model by the AIC. *Biometrika* 63(1): 117–126.
Shibata, R. 1980. Asymptotically efficient estimates of the order of a model for estimating parameters of a linear process. *Annals of Statistics* 8(5): 1147–1164.
Sims, C.A. 1972. Money income and causality. *American Economic Review* 62(4): 540–552.
Sims, C.A. 1980. Macroeconomics and reality. *Econometrica* 48(1): 1–47.
Slutsky, E. 1927. The summation of random causes as the source of cyclic processes. Trans. *Econometrica* 5: 105–146.
Tiao, G.C., and G.E.P. Box. 1981. Modeling multiple time series with applications. *Journal of the American Statistical Association* 76: 802–816.
Walker, G. 1931. On periodicity in series of related terms. *Proceedings of the Royal Society of London, Series A* 131: 518–532.
Wallis, K.F. 1977. Multiple time series analysis and the final form of econometric models. *Econometrica* 45(6): 1481–1497.
Wallis, K.F., and W.T. Chan. 1978. Multiple time series modeling: Another look at the mink–muskrat interaction. *Applied Statistics* 27(2): 168–175.
Whiteman, C.H. 1983. *Linear rational expectations models*. Minneapolis: University of Minnesota Press.
Whittle, P. 1983. *Prediction and regulation by linear least squares methods,* 2nd revised. Minneapolis: University of Minnesota Press.
Wilson, G.T. 1973. The estimation of parameters in multivariate time series models. *Journal of the Royal Statistical Society, Series B* 35(1): 76–85.
Wold, H. 1938. *A study in the analysis of stationary time series*. Stockholm: Almqvist and Wiksell.
Yamamoto, T. 1981. Prediction of multivariate autoregressive-moving average models. *Biometrika* 68(2): 485–492.
Yule, G.U. 1921. On the time-correlation problem with special reference to the variate-difference correlation method. *Journal of the Royal Statistical Society* 84: 497–526.
Yule, G.U. 1926. Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time series. *Journal of the Royal Statistical Society* 89: 1–64.
Yule, G.U. 1927. On a method for investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London, Series A* 226: 267–298.

Average Cost Pricing

K. J. Coutts

Average cost pricing and its associated variations refer to the practice of firms' price decisions. With the exception of certain markets for primary commodities and financial transactions where either an auctioneer or jobbers make prices, the vast majority of goods and services are traded in markets where firms must set prices. Once set, firms' sales are limited by the size of the market and the competition of rivals.

Following developments in the theory of the firm (Chamberlin 1933; Robinson 1933), interest

arose during the 1930s as to whether the pricing practices adopted by businesses provided supporting evidence for these theories. In their justly famous article of 1939, Hall and Hitch questioned 38 firms to discover what methods of price setting were actually applied and what motivated them to adjust prices. Their results revealed practices which appeared to be seriously at variance with the implications of received theory. Businesses typically set prices by calculating average costs of production and adding a mark up for profit. Firms did not habitually vary the mark up with variations in the strength of market demand. The findings were confirmed in other surveys after the War both in the UK (Andrews 1949) and in the detailed studies carried out in the USA (Kaplan et al. 1958).

According to these studies the precise method of price setting varied widely with firms and industries. In some cases the cost reference was average prime or variable costs. (Kalecki's degree of monopoly theory used this concept (Kalecki 1943).) In other cases fixed plus variable costs per unit of output or 'full cost' was common. Other variants reckoned unit costs at standard or normal levels of capacity utilization or of output. Depending therefore on the basis of unit cost adopted, the mark up might cover a target for gross profits alone, or would also include an allowance for fixed costs. While this type of price behaviour might be adopted by the industry leader, other firms might adopt a 'price minus' strategy of setting a target level of unit costs by deducting the firm's required mark up from the price set by the price leader or by foreign competition (Smyth 1967). Business interviews suggested that mark ups were relatively stable so that price changes moved mainly with changes in unit costs.

The evidence apparently conflicted with the theory. While it was conceded that the equality of marginal cost and revenue was an impractical operational procedure for setting price, mark up pricing adjustments could be interpreted as a useful 'rule of thumb' by which profits might be maximized by trial and error (Machlup 1946).

Econometric evidence has supplemented the original surveys. The literature is too large to summarize adequately here though Nordhaus (1971) provides a useful survey of US studies. An important aim of this literature was to use econometric methods to discover the extent to which the business cycle influenced the movement of prices. A disturbing feature of this work is the absence of uniform conclusions which can be made about price formation. There are no generally accepted measures of the relevant cost or demand variables and the results are often highly sensitive to the precise measures adopted. Some studies have generated a bewildering number of regressions correlating prices and indicators of costs and demand with little theoretical guidance or careful specification of hypotheses.

Although other economists (notably Kalecki 1943 and Gardiner Means 1935) have written about the concept of 'normal' or 'sticky' prices in the short run, Godley (1959) was the first to express the normal price hypothesis as a proposition about normal or standard unit costs. His hypothesis was that prices moved closely with normal unit costs and that the direct effect of demand on the mark up over normal unit costs was negligible. It assumed that firms operate with excess capacity and vary production principally by changing utilization rates. In forming price, firms are assumed to add a mark up to the average costs incurred when operating at a standard or normal rate of capacity utilization.

A series of empirical studies incorporated normal unit labour costs into price equations. Neild (1963) was the first to confirm support for this hypothesis using UK manufacturing data. Schultze and Tryon (1965), Fromm and Taubman (1968) and Eckstein and Fromm (1968) did similar studies for US data, though some of these found that capacity utilization measures of demand had an independent influence on price. Godley and Nordhaus (1972), in a study of the UK data, found that the effect of demand on prices was very small, once normal unit costs were measured appropriately in conformity with Godley's original hypothesis. Coutts et al. (1978) in a much larger study confirmed and extended the results to a number of sectors within manufacturing industry.

The empirical studies remain controversial, however, and have been subject to criticism.

Rushdy and Lund (1967), using Neild's data, publishing alternative specifications of the price equation in which significant demand effects appeared. Laidler and Parkin (1975) were critical of the tests of demand used by Godley and Nordhaus. McCallum (1970) demonstrated that manufacturing price changes correlated well with an indicator of excess demand alone, i.e. no explicit cost variables were included at all.

It is essential, in understanding the debate on the cyclical behaviour of costs and prices, to distinguish clearly variations in actual unit costs which arise as a consequence of a firm's own variations in capacity utilization from those which arise caused by factors outside the firm's immediate control. The former occur because in the short run productivity changes are dominated by changes in capacity utilization which impart a counter-cyclical movement to actual unit costs. This is partially offset by the pro-cyclical movement of wage earnings arising from variations in hours worked, piece rate bonuses and overtime payments. The latter type of cost variations may arise because of changes in negotiated wage rates, in materials and technology. Some of these may have a pro-cyclical character to the extent that the prices of basic commodities are sensitive to the business cycle and that wage rates increase more rapidly when labour markets are tight. This implies that while firms' actual unit costs are likely to be counter-cyclical, normal unit costs are more likely to be mildly pro-cyclical. The normal price hypothesis asserts that prices will therefore be as pro-cyclical as are normal unit costs. A direct implication is that the actual mark up will vary pro-cyclically and hence generate highly pro-cyclical variations in profits.

Given these cyclical properties the data would be consistent with all of the following: price is a non-cyclical mark up on normal unit costs; price is a pro-cyclical mark up on actual unit costs; price changes are directly proportional to excess demand alone. The empirical tests of the normal price hypothesis can claim to establish only that relative to normal unit costs, prices do not rise or fall with the course of the business cycle. If this evidence is accepted it implies that the putative influence of demand on price over the cycle is almost completely offset by the decline in actual, relative to normal, unit costs and hence that demand effects are probably small compared with costs in determining industrial prices. How then can theory accommodate this conclusion about the relative importance of costs and demand?

Studies of industrial cost characteristics have indicated that, within the relevant range of output variation, marginal costs are typically falling or flat rather than rising as might be the case in agriculture or mining (Johnston 1960). In the range where marginal costs are nearly constant they must also approximately equal average variable costs. By elementary manipulations of the profit maximizing conditions it follows that the optimal price may be expressed as a mark up on average variable cost – the mark up being a simple function of the firm's own elasticity of demand. This provides a common rationalization of the prevalence of mark up pricing practices in terms of the neoclassical theory of the firm. The interpretation apparently explains why costs have a major effect on prices while leaving a minor but significant role for demand (to alter the mark up) as theory predicts.

The difficulty in accepting this interpretation is that it does not explain the existence and persistence of underutilized capacity. It gives no convincing account why firms operate with a discretionary degree of spare capacity or why, in the absence of collusion, competition and profit maximizing behaviour does not force firms to operate at full capacity.

The requirement to set prices in industrial markets creates additional uncertainty for firms regarding expected market demand and the business strategies of current and potential rivals. Operating with a reserve capacity considerably increases the short-run flexibility of the firm to meet variations in demand, mainly by increasing hours of work and higher utilization of plant and machinery. Once prices are set, demand variations are first met by changing utilization rates.

By contrast prices perform a distinctive function in auction markets, conveying considerable information to buyers and sellers. Since the commodity can typically be classified into

homogeneous trading grades, it is unnecessary for a customer to know which seller produced the commodity purchased by the customer. Each is an anonymous participant in an auction market. The markets for most industrial products require instead that firms cultivate relations with their customers to encourage repeat sales. Prices convey only limited information about the characteristics of the product offered. Okun (1981) classified the latter as 'customer markets'. He argued that customer markets encourage the development of pricing policies of mutual benefit to producers and customers in which prices are largely determined by costs. The needs of producers to promote good-will makes them forgo any short run temporary advantage in raising price when demand strengthens. Customers are offered a stable price at which orders are placed unless costs of production change. Customer markets encourage product differentiation and non-price competition as methods of establishing a distinctive reputation with customers.

This tendency to cost-determined pricing may occur in markets where competitive pressure, as measured by the number of rival firms, is high. It is reinforced in oligopolistic markets where price changes that are unrelated to costs risk conveying signals to competitors which produce retaliatory responses. This observation underlies the kinked demand curve rationalization of Hall and Hitch (1939) and Sweezy (1939).

Mark up pricing implies that firms do not behave as if they were aiming to maximize profits in the short run, although they may have this objective among others over a longer time horizon. The accumulated empirical work on average cost pricing demonstrates the inadequacy of current microeconomic theory to explain how most industrial markets operate. It provides challenging material for economists to develop a richer theory of industrial competition.

See Also

- ▶ Administered Prices
- ▶ Marginal and Average Cost Pricing

References

Andrews, P.W.S. 1949. *Manufacturing business*. London: Macmillan.

Chamberlin, E.H. 1933. *The theory of monopolistic competition*. Cambridge, MA: Harvard University Press.

Coutts, K.J., W.A.H. Godley, and W.D. Nordhaus. 1978. *Industrial pricing in the United Kingdom*. Cambridge: Cambridge University Press.

Eckstein, O., and G. Fromm. 1968. The price equation. *American Economic Review* 58: 1159–1183.

Fromm, G., and P. Taubman. 1968. *Policy simulations with an econometric model*. Washington, DC: Brookings.

Godley, W.A.H. 1959. Costs, prices and demand in the short run. In *Macroeconomic themes*, ed. M.J.C. Surrey. Oxford: Oxford University Press. 1976.

Godley, W.A.H., and W.D. Nordhaus. 1972. Pricing in the trade cycle. *Economic Journal* 82: 853–882.

Hall, R.E., and C. Hitch. 1939. Price theory and business behaviour. *Oxford Economic Papers* 2: 12–45.

Johnston, J. 1960. *Statistical cost analysis*. New York: McGraw-Hill.

Kalecki, M. 1943. *Studies in economic dynamics*. London: Allen & Unwin.

Kaplan, A., J. Dirlam, and R. Lanzillotti. 1958. *Pricing in big business*. Washington, DC: Brookings.

Laidler, D., and M. Parkin. 1975. Inflation: A survey. *Economic Journal* 85: 741–809.

Machlup, F. 1946. Marginal analysis and empirical research. *American Economic Review* 36: 519–554.

McCallum, B.T. 1970. The effect of demand on prices in British manufacturing: Another view. *Review of Economic Studies* 37(1): 147–156.

Means, G.C. 1935. Price inflexibility and the requirements of a stabilizing monetary policy. *Journal of the American Statistical Association* 30: 401–413.

Neild, R.R. 1963. *Pricing and employment in the trade cycle*. Cambridge: Cambridge University Press.

Nordhaus, W.D. 1971. Recent developments in price dynamics. In *The econometrics of price determination*, ed. O. Eckstein. Washington, DC: Federal Reserve.

Okun, A. 1981. *Prices and quantities: A macroeconomic analysis*. Washington, DC: Brookings.

Robinson, J. 1933. *The economics of imperfect competition*. London: Macmillan.

Rushdy, F., and P.J. Lund. 1967. The effect of demand on prices in British manufacturing industry. *Review of Economic Studies* 34: 361–373.

Schultze, C.L. and J.L. Tryon. 1965. Prices and wages. In *The Brookings quarterly econometric model of the United States.* Chicago: Rand McNally.

Smyth, R. 1967. A price-minus theory of cost. *Scottish Journal of Political Economy* 110–117.

Sweezy, P.M. 1939. Demand under conditions of oligopoly. *Journal of Political Economy* 47: 568–573.

Averch–Johnson Effect

H. A. Averch

Abstract

The Averch–Johnson effect is produced when fair rate of return regulation encourages a firm to invest more than is consistent with the minimization of its costs. This can happen when the allowed rate of return exceeds the cost of capital, since the difference between the two represents pure profit. Detailed descriptions of actual regulatory processes may be useful in suggesting guides for action, since actual outcomes depend as much on political and bureaucratic necessity as they do on economic analysis and 'rational' benefit–cost estimates.

Keywords

Adjustment costs; Allocational efficiency; Averch, H. A.; Averch–Johnson effect; Bounded information; Bounded rationality; Capital–labour ratio; Cost of capital; Cost–benefit analysis; Electricity utilities; Innovation; Input–output analysis; Lagrange multiplier; Operational gaming; Production functions; Public utility pricing and finance; Rate of return; Rate of return regulation; Regulation, political economy of; Research and development; Simulation; Stochastic demand; Technical change; Technical efficiency

JEL Classifications

L5

The Averch–Johnson effect explores some unintended consequences of fair rate of return regulation (Averch and Johnson 1962). Such regulation may cause the firm to select excessively capital-intensive technologies, and, thereby, not produce its output at minimum social cost. Specifically, the main Averch–Johnson result is that the capital–labour ratio selected by a profit-maximizing, regulated firm will be greater than that consistent with a cost-minimizing one for any output it chooses to produce. If the fair rate of return is greater than the cost of capital, a firm will have an incentive to invest as much as it can consistent with its production possibilities, because the difference between the allowed rate and its actual cost of capital is pure profit.

This brief overview discusses (1) the effects of rate of return regulation on a monopolist's inputs and outputs; (2) the effects on incentives to innovate; (3) the empirical evidence on the existence and strength of the Averch–Johnson effect; and (4) some of the main theoretical extensions. Since 1962, the Averch–Johnson literature has been extended to include objectives other than profit maximization, more subtle interactions between regulators and firms and more complex market conditions. By making the models more complex, the number of possible regulatory outcomes has been enlarged. But the basic Averch–Johnson result, as stated above, has proven remarkably robust. So the discussion here focuses on this result and some of the main corollary results.

Choice of Inputs in the Basic Averch–Johnson Model

Suppose there exists a single-product, profit-maximizing monopolist subject to rate of return regulation. The firm's production function is

$$Q = F(K,L), K, L > 0, F(0,L) = F(K,0) = 0, \quad F_1, F_2 > 0, F_{11}, F_{22} < 0. \tag{1}$$

Suppose the firm's inverse demand function is

$$P = P(Q), P'(Q) < 0. \tag{2}$$

Profit is

$$\Pi = PQ - rK - wL. \tag{3}$$

Assuming, as is standard, that there is no depreciation and that the acquisition cost of

capital is adjusted to one, the rate of return constraint can be written

$$(PQ - wL)/K \leq s \text{ or } PQ - wL - sK \leq 0, \quad (4)$$

or

$$\Pi \leq (s - r)K, \quad (5)$$

where s is the allowed rate of return. The fair rate of return is taken to be at least as great as the cost of capital ($s > r$) and less than the rate the firm could earn if it were unconstrained. Consequently, the constraint is effective, and the firm maximizes

$$\Pi = PQ - rK - wL \quad (6)$$

subject to (4) or (5). Letting R equal total revenue PQ, the necessary first order conditions are

$$(1 - \lambda)R'F_1 - \lambda(s - r) = 0 \quad (7)$$

$$(1 - \lambda)R'F_2 - (1 - \lambda)w = 0 \quad (8)$$

$$R - wL - sK = 0. \quad (9)$$

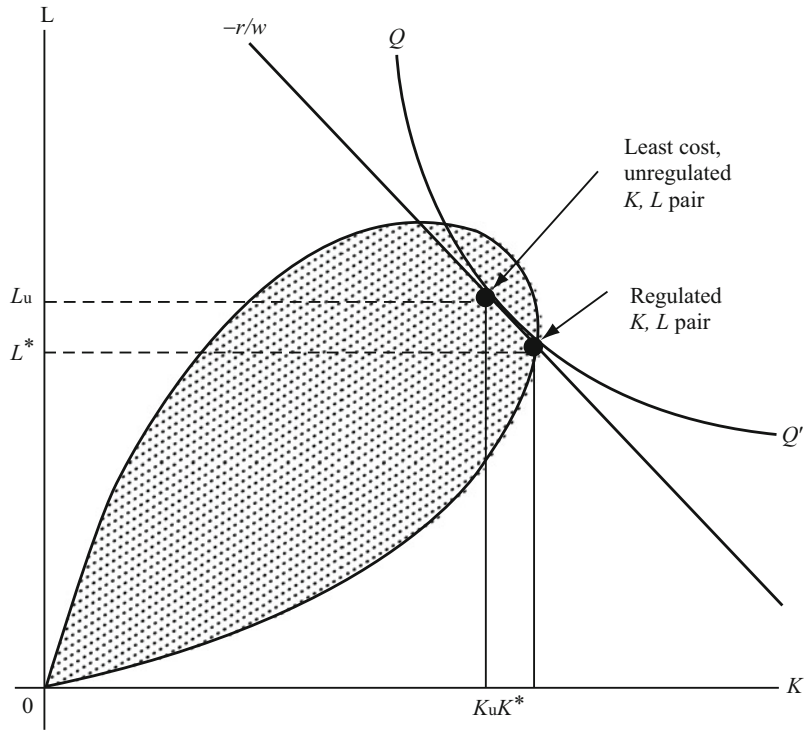
λ is the standard Averch–Johnson Lagrange multiplier. Given that the constraint is effective, that $s > r$, and that the revenue function $R = PQ$ is concave, the multiplier λ is greater than zero and less than one. Consequently, the marginal rate of substitution of capital for labour for the regulated firm is

$$-\mathrm{d}L/\mathrm{d}K = [r - (\lambda/1 - \lambda)(s - r)]/w < r/w \quad (10)$$

For any given output, the firm will not minimize cost, since this requires that the firm's marginal rate of technical substitution be equal to r/w.

This result can be shown graphically in several different ways (Baumol and Klevorick 1970; Zajac 1970). Zajac's formulation is shown here. Figure 1 shows the regulatory constraint (9) in relation to the firm's isoquants.

The shaded region inside the constraint curve shows input combinations resulting in rates of return greater than s. The firm wants to be as far up to the right on the constraint curve as possible, because, from (5), every increment of



Averch–Johnson Effect, Fig. 1 The Averch–Johnson (A-J) effect

capital increases profit. Consequently, the firm will operate at the rightmost point of the constraint curve.

The output for this rightmost point can be obtained from the isoquant that intersects the constraint curve at its rightmost point. However, the least cost combination of capital and labour for producing this output, where $-\mathrm{d}L/\mathrm{d}K = r/w$, lies inside the proscribed shaded area, on the firm's efficient expansion path. For any given output, the firm cannot simultaneously be on the cost-minimizing price line with slope $-r/w$ and on the constraint curve.

The Output of the Regulated Firm

One of the original rationales for regulation was that it would increase allocative efficiency by forcing monopolists to offer more output than ordinarily they would. If a larger output were always the result of rate of return regulation, then decreases in technical efficiency would be compensated by increases in allocational efficiency. In principle, regulatory agencies could seek an s that just balanced the marginal benefits of increased output against the marginal costs of decreased efficiency (Klevorick 1971; Sheshinski 1971; Bailey 1973; Callen et al. 1976).

Increasing output, however, is not inevitable. The firm will use greater quantities of capital as s falls towards r, but the amount of labour the firm chooses to use will not necessarily be larger, and so output need not be larger. However, if labour is not an inferior input – the most likely case – then the optimal amount of labour for a regulated monopoly will also increase over the unregulated one, and, consequently, so will output (Baumol and Klevorick 1970; Bailey 1973). Firms with linear, homogeneous production functions will produce greater output. Given two firms with identical positive, homogeneous production functions – one regulated, one unregulated – the output of the unregulated one becomes a lower bound on regulated output, and the output such that 'regulated' average cost equals price becomes an upper bound (Murphy and Soyster 1982).

Technological Change and the Regulated Firm

Even if regulated firms are inefficient in static situations, technological change conceivably could induce more output through cost reductions. And rate of return regulation might conceivably induce regulated firms to be more innovative than unregulated firms. Regulation usually guarantees some profits, if not maximal ones, and these could be used for innovation.

If technological change is exogenous to the firm, but is factor-augmenting, then the optimal constrained K^* rises (Westfield 1971; Magat 1976). However, factor- augmenting technological advance will not necessarily result in increased output, since the firm may again use less labour to produce its output. Technological change, of course, is not usually entirely exogenous. Through their own research and development (R&D), firms gain knowledge of feasible innovation possibilities. Profit-maximizing firms subject to both a rate of return constraint and their own innovation possibilities constraint can, depending on production conditions, choose more labour-augmenting technologies than they would without regulation, reinforcing the bias the regulated firm has towards relatively capital-intensive technologies (Smith 1974, 1975; Okuguchi 1975).

In any case, regulation does not unambiguously increase innovation possibilities. The R&D expenditures of the regulated firm are not always larger than those that an unregulated firm would select under the same production and demand conditions (Magat 1976). Furthermore, there is no systematic evidence that regulated firms select more high payoff R&D projects than unregulated ones and much anecdotal evidence to indicate that they are highly conservative.

Empirical Tests

In the mid-1970s and early 1980s there were a number of attempts to determine whether Averch–Johnson effects actually existed and whether, if they existed, they imposed significant social costs. The empirical investigations used different tests for the effect and different data sets, most, however, relating to electric utilities.

Unsurprisingly, the empirical evidence from these efforts was mixed. But overall the number

of empirical investigations that find some evidence for the Averch–Johnson effect or its behavioural consequences outnumber those that find no evidence.

Using different methods but similar data, Courville (1974) and Spann (1974) concluded that Averch–Johnson effects existed. Petersen (1975), using a costminimizing version of the Averch–Johnson model, found that as the allowed rate of return approached the market cost of capital, capital costs increased as did the share of those costs in total costs. Hayashi and Trapani (1976) confirmed that regulated firms have a capital–labour ratio greater than the cost-minimizing one and that tightening s decreases efficiency. However, Boyes (1976) concluded that there was no effect.

Smithson (1978) reported that there was static inefficiency among electric utilities, but he could not confirm that lowering the rate of return caused the optimal capital stock to increase. Tapon and Van der Weide (1979) found that only strictly regulated electric utility firms exhibit Averch–Johnson effects, but that less than half of the industry appears to be so regulated. Regulatory lag permits firms to avoid Averch–Johnson effects, but raises the question of the worth of public investments in regulatory institutions.

Gollop and Karlson (1980), using data on electric utilities and an intertemporal model, found no evidence of input distortions. But Filer and Hallas (1983), testing for the effects of regulation in the interruptable gas industry, found rate of return regulation induced investment in additional storage capacity. Giordano (1983), examining utilities during 1964–77, concluded that there was capital bias during the 1960s, but not in the 1970s, because increasing regulatory lag and rapidly rising factor prices wiped it out. Such a finding was consistent with Averch–Johnson predictions, but it made Averch–Johnson effects perhaps less relevant in the 1980s. However, Averch–Johnson effects continue to be reported. Mirucki (1984), for example, concludes that the Canadian Bell system overinvests in capital and does not minimize costs.

Some investigators have argued that even if Averch–Johnson effects exist, their impact may be small, for there may be deterrents to technical inefficiency such as open entry (Sharkey 1982). Others have argued that even if Averch–Johnson effects existed in the 1960s and 1970s, the relevant problem for utilities in the 1980s has been one of avoiding actual rates of return that fall below the allowed rate s. The 1980s problem is under-investment, because consumers are now able to prevent regulatory agencies from granting the price increases necessary to cover rising input costs (Navarro 1983; Nelson 1984; Rozek 1984).

Theoretical Extensions

The Averch–Johnson results have been extended and generalized in many ways. Three of the more significant extensions are discussed below.

Regulatory lag and stochastic review: The original Averch–Johnson result implicitly assumed regulatory agencies were always effective in enforcing the s they chose. In fact, regulators have great difficulty in keeping actual rates close to target rates. The regulatory process does its work episodically, through adjustments in price. Occasional adjustments, the regulator hopes, will bring the actual s to a tolerable level, if not back to the one originally set. Since regulation is a political, bureaucratic and legal process, there are almost always lags in enforcement. Consequently, firms may be able to escape the constraint for long periods of time (Bailey and Coleman 1971; Klevorick 1973).

Sufficient regulatory lag may allow the firm to be technically efficient at an unregulated monopolist's output, and it may induce more technological innovation than the case without enforcement lag. Continuous, effective regulation would prevent the firm from gaining the windfall profits that innovation may require, although Nelson argues that most technological change in the utilities industry is disembodied and has little relation to regulation (Nelson 1984).

Demand uncertainty: Some authors argue that Averch–Johnson results hold only under some specifications of a stochastic demand function, but not others (Perrakis 1976; Peles and Stein 1976). Most of this discussion goes to whether the optimal capital stock would be larger, if regulated firms faced stochastic demands. If, as in the original Averch–Johnson discussion, we assume that the firm selects K and L as part of a

simultaneous, ex ante optimization process, then the basic Averch–Johnson result, the inefficient capital–labour ratio, still holds under stochastic demand (Das 1980).

Dynamic analysis: Some authors have introduced time explicitly into the original static Averch–Johnson model. For example, El-Hodiri and Takayama (1981) interpret the 'Averch–Johnson effect' to be a larger optimal K^* for a regulated firm than an unregulated one, and they show that this is true even with the adjustment costs attributable to time. However, much of this dynamic literature has been devoted to showing that, given a firm that maximizes the present value of profits over any number of time periods, one or more Averch–Johnson results do not hold or hold only under special conditions (Niho and Musaccio 1983; Dechert 1984).

The Significance of the Averch–Johnson Effect

From the stand-point of microeconomic theory, the original Averch–Johnson results provided impetus for increasingly complex, analytical models of the regulatory process. The Averch–Johnson approach suggested that much of the conventional, qualitative wisdom about regulation could be modelled and tested and that it was necessary to do so. Without thinking through all the potential consequences, actions and rules could be quite flawed without anyone intending them to be so. But flaws generally become apparent only after actions and rules have become entrenched, difficult to change or reverse. So explicit modelling of regulatory rules became part of the economist's stock in trade.

From a public policy perspective, the Averch–Johnson results and the very large volume of follow-on research have made economists, legislators and administrators far more sensitive to the potential unintended consequences of regulatory alternatives in general and not just rate of return alternatives. The Averch–Johnson effect has also figured directly in rate cases with utilities sometimes forced to defend themselves against charges of inefficiency.

Future Lines of Development

By injecting changes into the Averch–Johnson formulation one at a time, theoretical work has sought to make the model more representative of the actual regulatory process. One set of writers has pursued the effects of stochastic demand. Another set has worked on regulatory lag and stochastic review processes, but without stochastic demand. Yet another set has had the firm making global optimizations over time without either stochastic demands or random review. Economists interested in welfare issues have tried to determine an optimal fair rate of return from a strict economics perspective, but neglected politics and bureaucratic behaviour in setting rates. No model builders to date have addressed firms and regulators as interacting organizations both suffering from bounded rationality and bounded information, although there is some recent work on what regulators might do when a firm's costs are unknown and it has incentives to lie (Baron and Meyerson 1982).

Regulatory systems are so complex and interactive that the standard strategy of a priori modelling with a minimum number of plausible assumptions may no longer have sufficient pay off. In complex, interactive, relatively poorly understood situations, other analytical styles such as simulation or operational gaming can be useful. They have not been tried and probably should be. In fact, brute force, detailed descriptions of actual regulatory processes may be highly useful in suggesting guides for action. Regulation remains a problem in political economy. Actual outcomes depend as much on political and bureaucratic necessity as they do on economic analysis and 'rational' benefit–cost estimates.

See Also

► Marginal and Average Cost Pricing

Bibliography

Averch, H.A., and L.L. Johnson. 1962. Behavior of the firm under regulatory constraint. *American Economic Review* 52: 1052–1069.

Bailey, E.E. 1973. *Economic theory of regulatory constraint*. New York: D.C. Health.

Bailey, E.E., and R.D. Coleman. 1971. The effect of lagged regulation in an Averch–Johnson model. *Bell Journal of Economics and Management Science* 2 (1): 278–292.
Baron, D.P., and R.B. Meyerson. 1982. Regulating a monopolist with unknown costs. *Econometrica* 50 (4): 911–930.
Baumol, W.J., and A.K. Klevorick. 1970. Input choices and rate-of-return regulation. *Bell Journal of Economics and Management Science* 1 (2): 162–190.
Boyes, W.J. 1976. An empirical examination of the Averch–Johnson effect. *Economic Inquiry* 14 (1): 25–35.
Callen, J., G.F. Mathewson, and H. Mohring. 1976. The benefits and costs of rate of return regulation. *American Economic Review* 66 (3): 290–297.
Courville, L. 1974. Regulation and efficiency in the electric utility industry. *Bell Journal of Economics and Management Science* 5 (1): 53–74.
Das, S.P. 1980. On the effect of rate of return regulation under uncertainty. *American Economic Review* 70: 456–460.
Dechert, W. 1984. Has the Averch–Johnson effect been theoretically justified? *Journal of Economic Dynamics and Control* 8 (1): 1–17.
El-Hodiri, M., and A. Takayama. 1981. Dynamic behavior of the firm with adjustment costs under regulatory constraint. *Journal of Economic Dynamics and Control* 3 (1): 29–41.
Filer, J.E., and D.R. Hallas. 1983. Empirical tests for the effect of regulation on firms and interruptible gas service. *Southern Economic Journal* 50 (1): 195–205.
Giordano, J.N. 1983. The changing impact of regulation on the U.S. electric utility industry, 1964–1977. *Eastern Economic Journal* 9 (2): 91–101.
Gollop, F.M., and S.H. Karlson. 1980. The electric power industry: An econometric model of intertemporal behavior. *Land Economics* 56 (3): 299–314.
Hayashi, P.M., and J.M. Trapani. 1976. Empirical evidence on the Averch–Johnson model. *Southern Economic Journal* 42 (3): 384–398.
Klevorick, A.K. 1971. The 'optimal' fair rate of return. *Bell Journal of Economics and Management Science* 2 (1): 122–153.
Klevorick, A.K. 1973. The behavior of a firm subject to stochastic regulatory review. *Bell Journal of Economics and Management Science* 4 (1): 57–88.
Magat, W. 1976. Regulation and the rate of direction of induced technical change. *Bell Journal of Economics and Management Science* 7 (2): 478–496.
Mirucki, J. 1984. A study of the Averch–Johnson effect in the telecommunications industry. *Atlantic Economic Journal* 12 (1): 121.
Murphy, F.H., and A.L. Soyster. 1982. Optimal output of the Averch–Johnson model. *Atlantic Economic Journal* 10 (4): 77–81.
Navarro, P. 1983. Save now, freeze later: The real price of cheap electricity. *Regulation* 7: 31–36.
Nelson, R.A. 1984. Regulation, capital vintage, and technical change in the electric utility industry. *Review of Economics and Statistics* 66 (1): 59–69.
Niho, Y., and R.A. Musaccio. 1983. Effects of regulation and capital market imperfection on the dynamic behavior of a firm. *Southern Economic Journal* 49: 625–636.
Okuguchi, K. 1975. The implications of regulation for induced technical change: Comment. *Bell Journal of Economics and Management Science* 6: 703–705.
Peles, Y.C., and J.L. Stein. 1976. The effect of rate of return regulation is highly sensitive to the nature of uncertainty. *American Economic Review* 66: 278–289.
Perrakis, S. 1976. On the regulated price-setting monopoly firm with a random demand curve. *American Economic Review* 66: 410–416.
Petersen, H.C. 1975. An empirical test of regulatory effects. *Bell Journal of Economics and Management Science* 6 (1): 111–126.
Rozek, R.P. 1984. The over-capitalization effect with diversification and cross subsidization. *Economics Letters* 6 (1–2): 159–163.
Sharkey, W.W. 1982. *The theory of natural monopoly.* New York: Cambridge University Press.
Sheshinski, E. 1971. Welfare aspects of a regulatory constraint: Note. *American Economic Review* 61: 175–178.
Smith, V.K. 1974. The implications of regulation for induced technical change. *Bell Journal of Economics and Management Science* 5: 623–632.
Smith, V.K. 1975. The implications of regulation for induced technical change: Reply. *Bell Journal of Economics and Management Science* 6: 706–707.
Smithson, C.W. 1978. The degree of regulation and the monopoly firm: Further empirical evidence. *Southern Economic Journal* 44: 568–580.
Spann, R.M. 1974. Rate of return regulation and efficiency in production: An empirical test of Averch–Johnson thesis. *Bell Journal of Economics and Management Science* 5 (1): 38–52.
Tapon, F., and J. Van der Weide. 1979. Effectiveness of regulation in the electric utility industry. *Journal of Economics and Business* 31 (3): 180–189.
Westfield, F. 1971. Innovation and monopoly regulation. In *Technological change in regulated industries*, ed. W.M. Capron. Washington, DC: Brookings Institution.
Zajac, E.E. 1970. A geometric treatment of Averch–Johnson's behavior of the firm model. *American Economic Review* 60: 117–125.

Axiomatic Theories

Patrick Suppes

One of the first steps in axiomatizing a theory is to list the primitive notions. A familiar example is

the classical case of Euclidean geometry. We can take as primitives the following three notions: the notion of point, the notion of betweenness – one point being between two others in a line – and the notion of equidistance – (the distance between given points being the same as the distance between two other given points). Other geometric notions can then be defined in terms of these three notions. For example, the line generated by two distinct points a and b is defined as the *set* of all points c which are between a and b, which are such that b is between a and c, or which are such that a is between c and b.

The primitive notions of a theory are seldom, if ever, uniquely determined by the intuitive content of the theory. Euclidean geometry has been developed in terms of a wide variety of primitive notions other than the three mentioned above. In Hilbert's well-known axiomatization (1899), for example, the five notions of point, line, plane, betweenness and congruence are taken as primitive. In contrast, in the same year, the Italian mathematician Pieri published an axiomatization of Euclidean geometry using only the primitive notions of point and motion.

An important preliminary step in fixing on the primitive notions of a theory is to make explicit what other theories are to be assumed in developing the axiomatization. For most axiomatic work in economics, a certain amount of standard mathematics is assumed, including of course logic and elementary set theory, as well as most of classical analysis. When such prior theories are not assumed, then a complete apparatus must be built from the ground up. This is quite uncommon in any of the empirical sciences, such as economics or physics. For example, it would seem strange in a theoretical paper in economics to develop from scratch the concept of number or the concept of Riemann integral.

After making explicit what other theories are to be assumed and fixing on the primitive notions of the theory under study, the axioms of the theory can now be stated without ambiguity. The only concepts referred to in the axioms must be primitive notions, notions defined in terms of primitive notions, or notions belonging to the theories assumed a priori. It is also important to recognize that in deriving theorems of the theory in question, nothing may be assumed about the primitive notions except what is stated in the axioms or possibly follows from other theories assumed a priori.

Informally, there are other things that are often said about axioms which can be repeated here but which cannot always be satisfied. For instance, it is generally recognized that it is desirable to have as few axioms as feasible, and also to take as axioms statements which have a strong intuitive appeal. But minimization of number of axioms or the vague concept of intuitive appeal do not play explicitly a rigorous role in almost any critiques of actual axiom systems proposed. It is sometimes held that theories should always be formulated only in terms of their primitive notions; that is, without using any notions defined in terms of the primitives. The argument for this is that only in this fashion will the actual complexity of the theory be evident. Already in the case of axioms of geometry this can become an intolerable burden from the standpoint of perspicuity of formulation, and consequently it is again a recommendation that is to be followed when feasible but not taken as an inviolable injunction.

The rest of this article is organized in the following fashion. Section "History" provides a brief review of the history of the axiomatic method. Section "Theories with Standard Formalization" analyses the concept of the standard formalization of a theory in first-order logic, and points out why this approach does not work well in most scientific contexts. The positive approach of considering theories as being defined by set-theoretical predicates is developed in section "Theories Defined as Set-Theoretical Predicates".

History

Euclid

As with many other things, the story of the axiomatic method begins with the ancient Greeks. It seems fairly certain that it developed in response to the early crisis in foundations; namely, the problem of incommensurable magnitudes as, for instance, of the side and diagonal of a square,

which occurred in the fifth century BC. The axiomatic method as we think of it today was crystallized in Euclid's *Elements*. The important philosophical predecessor of Euclid is Aristotle, who discusses the first principles of any demonstrative science in the *Posterior Analytics*. According to Aristotle, a demonstrative science must start from indemonstrable principles. Of such principles Aristotle says in the *Posterior Analytics* that some are common to all sciences. These are what are termed *axioms* or *common notions*. Other principles are special to a particular science. A standard example of an axiom for Aristotle is the principle that if equals be subtracted from equals, the remainders are equal.

Euclid follows Aristotelian methodology by listing at the beginning of the *Elements* 23 definitions, 5 postulates (assumptions special to geometry) and 5 axioms or common notions. Euclid set the standard of rigour for nearly 2,000 years of mathematics. Only in the nineteenth century did real flaws come to the surface in Euclid's axiomatic presentation. It should also be mentioned that in spite of the usefulness of the ancient distinction between postulates and axioms, it is not one that is explicitly made today. What Euclid called axioms are now taken up in what has been referred to above as the theories that are assumed a priori; for example, logic and classical mathematics. What are called axioms in the context of this article would be called postulates in ancient Greek terminology, but this ancient usage will not be followed here.

Although from a modern standpoint it is easy to pick out certain flaws in Euclid's *Elements* and to emphasize certain differences between his conception of the axiomatic method and modern ones, the essential point remains that the axiomatic method as reflected in his *Elements* is extremely close to modern views. Such important works of modern science as Newton's *Principia* (1687) were written in this geometrical tradition.

Modern Geometry

The historical source of the modern viewpoint towards the axiomatic method was the intense scrutiny of the foundations of geometry in the nineteenth century. The most important driving force behind this effort was the discovery and development of non-Euclidean geometry at the beginning of the nineteenth century by Bolya, Lobachevski and Gauss.

It was above all the German geometer Pasch who formulated in the clearest and most explicit way the modern formal conception of geometry in his important book of 1882. Pasch emphasized that if geometry is to be a genuinely deductive science, then the deductions must everywhere be independent of the meaning of geometrical concepts. From a formal standpoint, the deductions should be valid without taking into account in any way the meaning of the terms. He emphasized the thoroughly modern point that if a theorem is rigorously derived from a set of axioms, and if we replace the primitive concepts by others of the same logical nature, then the theorem will remain valid. This has the effect of treating the primitive concepts as variables. The axiomatic approach in geometry continued to dominate conceptions of the axiomatic method well into the twentieth century.

Theories with Standard Formalization

The most explicit and formally precise axiomatic versions of theories are ones that are formalized within first-order logic with identity. First-order logic can be easily characterized in an informal way. This is the logic that assumes (i) one kind of variable; (ii) logical constants, in particular the sentential connectives such as $\rightarrow$ for *if . . . then . . .*, and v for *or*; (iii) the universal and existential quantifiers, (x), $(\exists x)$; and (iv) the identity symbol. A theory formulated within such a framework is called a theory with standard formalization. Three kinds of non-logical constants occur in axiomatizing the theory: the predicates or relation symbols, the operation symbols and the individual constants.

The expressions of the theory – i.e. finite sequences of symbols, of the language of the theory – are divided into terms and formulas. Recursive definitions of each are given. The simplest terms are variables or individual constants. New terms are built up by combining simpler

terms with operation symbols in the appropriate fashion. Atomic formulas consist of a single predicate and the appropriate number of terms. Compound formulas are built up from atomic formulas by means of sentential connectives and quantifiers.

Theories with standard formalization are not often found in use in scientific context. They do have a role when particular questions are of interest. For example, a standard question about a theory with standard formalization is whether it is decidable. This means, is there a mechanical decision procedure for asserting whether or not a formula of the theory is a valid sentence of the theory, i.e. is a formula either an axiom or a theorem of the theory? In general, there is no decision procedure for theories with standard formalization. It was rigorously proved in 1936 by Alonzo Church that there is no mechanical test for the validity of arbitrary formulas in first-order logic. The most important positive decision result is probably that of Alfred Tarski's for the elementary algebra of real numbers, first published in 1948. A second important question is that of completeness. It is natural to ask of any scientific theory whether it is complete in the sense that it is possible to give a list of axioms of the theory from which all other true assertions of the theory may be derived. The most important result of modern logic is Kurt Gödel's result in 1931 that the elementary theory of positive integers is not complete in the sense just stated.

There are difficulties of casting ordinary scientific theories into first-order logic. The source of the difficulty has already been mentioned. Almost all systematic scientific theories assume a certain amount of mathematics a priori. There is no simple or elegant way to include such mathematical concepts in the standard formalization, which by definition assumes only the apparatus of elementary logic. For example, a theory that requires for its formulation an Archimedean-type axiom – for some n, n copies of a length however small are together longer than any given distance no matter how long – cannot be axiomatized in first-order logic. Because of these difficulties, standard axiomatic formulation of scientific theories follows the methodology outlined in the next section.

Theories Defined as Set-Theoretical Predicates

From a formal standpoint, the essence of the approach that is close to the practice of modern mathematics and widely used in mathematical economics is to axiomatize scientific theories within a set-theoretical framework. From this standpoint, to axiomatize a theory is simply to define a certain set-theoretical predicate. The axioms as we ordinarily think of them are part of a definition, of course the most important part.

Here is the theory of weak orderings formulated as such a set-theoretical definition – for an elementary exposition of this approach, see Suppes (1957, chapter 12).

Definition Let A be a non-empty set and R a binary relation on A. A structure (A,R) is a *weak ordering* if and only if for every x, y, and z in A (i) if xRy and yRz then xRz, (ii) xRy or yRx.

Further formal work is then conducted in terms of the structures of a theory as thus defined. For many kinds of analysis a key definition is that of isomorphism of structures, here exemplified for the structures just considered.

Definition A structure (A,R) is isomorphic to a structure (A',R') if and only if there is a function f such that (i) the domain of f is A and the range of f is A', (ii) f is a one–one function, (iii) if x and y are in A, then xRy if and only if $f(x)R'f(y)$.

In terms of the definition of isomorphism, we then often seek for axiomatic theories a representation theorem which has the following meaning. A certain class of structures or models of a theory is distinguished for some intuitively clear reason and is shown to exemplify within isomorphism every structure of the theory. In the case of ordering relations, a typical representation theorem concerns representing any ordering by an isomorphic numerical ordering. Note that in the case of weak orderings, for example, we first form classes of objects that are equivalent in the ordering and then each equivalence class is assigned a number under the isomorphic representation.

Only the simplest kinds of structures have been discussed here, but the ideas developed apply

without change to more complicated structures as exemplified in contemporary theories of pure mathematics, mathematical economics and other mathematically based disciplines. The development of axiomatic theories for such structures is now the widely accepted methodology for their investigation. The importance of identifying certain structures as basic is perhaps obvious, but a persuasive explicit argument is given in Bourbaki (1950) for the case of mathematics.

Finally, it should be emphasized that the commitment to a set-theoretical framework for the formulation of axiomatic theories is less essential than the general *formal* conception of the axiomatic method as it originated in modern geometry since Pasch and has been developed in detail in pure mathematics by Bourbaki.

See Also

- ▶ Orderings
- ▶ Philosophy and Economics
- ▶ Preferences
- ▶ Preordering

References

Bourbaki, N. 1950. The architecture of mathematics. *American Mathematical Monthly* 57: 231–232.
Euclid. *Elements*. Heath translation, 2nd ed., 1925; reprinted, New York: Dover, 1956.
Hilbert, D. 1899. *Gründlagen der Geometrie*, 9th ed. Stuttgart. 1962.
Pasch, M. 1882. *Vorlesungen über neuere Geometrie*. Leipzig: Springer.
Suppes, P. 1957. *Introduction to logic*. New York: Van Nostrand.

Ayres, Clarence Edwin (1891–1972)

Warren J. Samuels

Keywords

Ayres, C. E.; Culture; Industrialization; Institutional economics; Methodological collectivism; Technology; Values; Veblen, T.

JEL Classifications

B31

Ayres was born on 6 May 1891 in Lowell, Massachusetts, and died on 25 July 1972 in Alamogordo, New Mexico. Trained as a philosopher, with degrees from Brown and Chicago (PhD, 1917), Ayres taught at Chicago, Amherst and Reed before moving to the University of Texas at Austin in 1930, from which he retired in 1968. For one year, 1924–5, he was an associate editor of *The New Republic,* associated with Herbert Croly, John Dewey, Alvin Johnson and R.H. Tawney. He had a lifelong correspondence with another philosophically oriented, but more traditional economist, Frank H. Knight.

He was profoundly influenced by Thorstein Veblen and Dewey and became a, if not the, leader of institutional economics after World War II. A truly charismatic lecturer, at Texas he had long-lasting influence on a coterie of students who continued his teachings in their own careers. As his ideas evolved, particularly with regard to the nature of and relations between institutions and technology, his students came away with coherent but varying substantive understandings.

Ayres' formulation of institutionalism stressed that science was a system of belief, that human values were only means to the continuation and enhancement of the life process, that technology, as he defined it, was a (largely) beneficent driving force in social change, and that considerations of rightness tended in practice to be matters of tradition and custom.

Technology, to Ayres, meant the use of tools, but he defined tools increasingly broadly to include intangible symbols and organizations. Technology was the surging force governing economic welfare, and constituted what he considered to be an objective industrial or developmental process. His conception of technologically instrumental value and truth emphasized the transcultural values of workability and efficiency which form a continuum. Opposed to technology was the binding force of established institutions which, through sanctioning ceremonial behaviour in favour of established or vested interests, were hostile to the conceptual and economic progress

generated by technology. Economic progress was thus fundamentally a matter of industrialization; the logic of industrialization, or technological advancement in all respects, was continually at war with outworn, inhibitive institutions. Mankind's task was to develop new institutional forms and revise old ones in order to keep pace with evolving technology.

Ayres insisted that human behaviour was socially formed, and that for such behaviour to be explained and understood the economist had to study existing behaviour patterns (institutions) and general culture. In common with other institutionalists, Ayres insisted upon methodological collectivism and challenged what he considered to be the narrow focus on market equilibrium conditions maintained by mainstream economics.

Ayres influenced many development economists, who similarly perceived that modernization was inhibited by the continuance of traditional institutions or by the maintenance of positions of power antagonistic to modernization. More generally, Ayres, again like other institutionalists, argued that to understand the allocation of resources one had to go beyond the market to the institutions and cultural forces which, in part through adaptation to and incorporation of technology, constitute the real allocational mechanism. In a sense, the neoclassical juxtaposition between cost of production and utility became for Ayres something different, a juxtaposition between technology and the institutions which formed and weighted individual and collective choice.

See Also

► Institutional Economics

Selected Works

1938. *The problem of economic order.* New York: Farrar & Rinehart.
1944. *The theory of economic progress.* Chapel Hill: University of North Carolina Press.
1946. *The divine right of capital.* Boston: Houghton Mifflin.
1952. *The industrial economy.* Boston: Houghton Mifflin.
1961. *Toward a reasonable society.* Austin: University of Texas Press.

Bibliography

Breit, W. 1973. The development of Clarence Ayres's theoretical institutionalism. *Social Science Quarterly* 54: 244–257.