F

Fabian Economics

Elizabeth Durbin

Keywords

Beveridge, W. H.; Capitalism; Clark, C. G.; Cole, G. D. H.; Dalton, H.; Durbin, E.F. M.; Equality; Exchange value; Fabian economics; Gaitskell, H. T. N.; Imperfect competition; Individual freedom; Inequality; Inheritance; Keynesian Revolution; Lewis, W. A.; Marginal productivity theory; Marginal revolution; Market socialism; Meade, J. E.; Mixed economy; Monopoly capitalism; Nationalization; Planning; Property; Public ownership; Redistribution of income and wealth; Shaw, G. B.; Social democracy; Socialism; Tawney, R. H.; Trade cycle; Webb, B. and S

JEL Classifications

O1

Despite the Webbs' disdain for abstract economics ('sheer waste of time'), economic arguments have always held a central place in the Fabian case for socialism. As in most matters the Fabian Society has approached the dismal science eclectically. Some members have accepted market economics, others have rejected it; some embraced the Keynesian Revolution, others remained sceptical; some have believed in market pricing, others have been convinced that controls are essential for centralized planning. There is no consistent body of thought which could properly be described as Fabian economics. There is nonetheless a distinctive Fabian approach to economics, which this essay identifies while tracing the significant shifts in its key elements.

The Fabians and the Marginal Revolution

When the small group including Sydney Olivier, Bernard Shaw and Sidney Webb first started to meet at Mrs Charlotte Wilson's house in Hampstead, they set themselves the task of reading Marx's *Das Kapital* chapter by chapter. Graham Wallas, who joined the group in February 1885, later recalled how they were astonished to find 'that we did not believe in Karl Marx at all' (Wallas 1923). Webb, Wallas and Shaw were also members of the Economic Circle, an offshoot of the Bedford Chapel Debating Society, where Professor Edgeworth helped to expound the principles of the new marginal economics with another economist, Philip Wicksteed. Thus, according to Wallas, under Webb's leadership the group thrashed out 'the Jevonian anti-Marx value theory as the basis of our socialism'. Shaw apparently needed more convincing than others. In the *Fabian Essays* he later described how he had been converted from his earlier Marxist faith

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

that the working class revolution would take place in Britain by 1889 'at latest' (Shaw 1908, pp. 218–19). Instead of manning the barricades that year, Shaw was busily explicating the new Fabian economic basis for socialism.

In his preface to the essays, Shaw explained that the writers were all social democrats, 'with a common conviction of the necessity of vesting the organisation of industry and the material of production in a State identified with the whole people by complete Democracy'. In his contributions he propounded the theory of marginal productivity, demonstrating that Ricardian economic rent, or 'surplus value', can accrue to all the factors of production, to land and to labour, and not just to capital as in the Marxist version. Similarly, he rejected the labour theory of value, and advanced the neoclassical version, which he called 'exchange value'; in other words value was determined by the interaction of supply and demand in the marketplace. Shaw concluded:

> What the achievement of Socialism involves economically, is the transfer of rent from the class which now appropriates it to the whole people. Rent being that part of the produce which is individually unearned, this is the only equitable method of disposing of it. (1908, p. 220)

The method proposed to accomplish the transition was the common ownership of property, or as Webb put it: 'the gradual substitution of organized operation for the anarchy of competitive struggle' (p. 62).

The original essayists all shared Marx's moral outrage at the evils of capitalism, particularly as a cause of hopeless poverty, inhuman working conditions and excessive inequality, and they also identified the institution of private property as its prime motivating force. However, they did not share the Marxist belief that capitalism must inevitably collapse. Although they recognized that periodic slumps were endemic to the system, they were more struck by its spectacular long-run growth and saw no reason to suppose that it would not continue to reap the benefits of technological change. Thus, as Schumpeter later explained, they were the kind of socialists who believed in the productive success of capitalism while they deplored its distributive results (Schumpeter 1942, pp. 61–2). They thought that through the gradual extension of public property socialism would evolve from democratic efforts to mitigate the effects of industrialization. Indeed, Webb provided an extraordinary two-page catalogue of socialism's accomplishments to date, which ranged from the army and navy to public baths and cow meadows (Shaw 1908, pp. 66–7). William Clarke described the growth of joint stock companies, and more recently of 'rings' and 'trusts', through which ownership became ever more divorced from entrepreneurial function and 'capitalism ever more inconsistent with democracy and the public interest'. These changes provided the other main Fabian justifications for the public ownership of industry.

Their views on the actual operations of a socialist system were hazy. Shaw and Webb both imply that socialism will have arrived when the entire market operation is administered through nationalization, municipalization and government regulation. Shaw described the aim of social democracy:

> to gather the whole people into the state, so that the state may be trusted with the rent of the country, and finally with the land, the capital, and the organisation of the national industry—with all sources of production, in short, which are now abandoned to the cupidity of irresponsible private individuals. (1908, p. 224)

Yet, in other Fabian tracts, Shaw extolled the virtues of competition and of individual freedom, asserting that the latter was 'as highly valued by the Fabian Society as Freedom of Speech, Freedom of Press, or any other article in the charter of popular liberties' (Shaw 1896, p. 327).

Later, of course, the Webbs provided a far more detailed view of their ideas for the organization of a Social Parliament to decide economic policy and to administer public enterprises. Beatrice herself remained ambivalent as to whether unemployment was caused by personal failings or 'the disease of industry'; their apparently countercyclical unemployment scheme only shifted existing projects without requiring fundamental changes in government policy (Harris 1972, pp. 42–3). The Webbs' ideas about state planning were based on administrative principles, not economic science.

In the next Fabian generation, Hugh Dalton, a student of Pigou, used Pigou's revised version of

neoclassical theory to demonstrate the critical differences between factor incomes and personal incomes. He introduced and defined the nature of inheritance and its role in maintaining wealth differentials; he broadened its concept to include educational opportunity, access to public services and institutional customs (Dalton 1920). According to Gaitskell's later assessment of the British tradition, Dalton's work was a decisive influence in shifting socialist thought from the 'sterile, out-of-date, somewhat academic arguments of earlier writers' to the practical issues of progressive taxation and educational reform (Gaitskell 1955, pp. 936–7).

Although still grounded in neoclassical criteria of allocative efficiency, Dalton's analysis dealt directly with income equality, opening up ways to achieve socialism other than through Webbian public ownership. Thus, Gaitskell believed that the case for socialist equality could be stated on 'straightforward ethical principles', rather than on 'complicated arguments about economic abstractions'.

The Fabians, the Keynesian Revolution and Economic Planning in the 1930s

The Great Depression threatened both the political and economic stability of capitalist systems. Inspired by the Russian Revolution and its apparent success in replacing capitalism and avoiding mass unemployment, many leftist sympathisers turned to Marxism. They struggled through *Das Kapital*, they visited the Soviet Union, and they recommended the Soviet political philosophy and economic system. The Webbs fell in love with Russia; in their last major work, *Soviet Communism*: *A New Civilization*?, they advocated a totally controlled economy, visualising Soviet planning as the ultimate Fabian collective. In *New Fabian Essays* Crossman argued that they had simply superimposed Marxism on their basic utilitarianism; he believed that only John Strachey successfully re-thought the entire system 'in Anglo-Saxon terms' (Crossman 1970, p. 5).

It fell to the younger generation to restate the traditional Fabian case against Marxist economic thought and revolutionary methods and to redefine the democratic socialist alternative. Hugh Gaitskell and Evan Durbin organized the Economic Section of the New Fabian Research Bureau, which had been founded by G.D.H. Cole in March 1931 and merged with the parent Fabians in 1938; their purpose was to explore the implications of the theoretical economic controversies for socialism and to make policy recommendations to the Labour Party (Durbin 1985). At the same time the obvious failures of the market system were challenging economists to rethink the role of government intervention and to redesign their toolkit. Keynesian macroeconomics, the economics of imperfect competition and the principles of economic planning embodied in the new 'market socialism' were first developed during the 1930s. After the war they were incorporated into the orthodox case for the mixed economy.

In pointed contrast to official policy, Keynes had begun pressing British governments to expand, not to contract, public expenditure to cope with unemployment. In the early 1930s his position was largely intuitive; *The General Theory* published in 1936 was the first systematic exposition of his theoretical case. Until then the most fundamental cleavage on the unemployment issue was between those who advocated government intervention in the market and those who did not. Socialists were naturally allied with the interventionists on social and political grounds, as well as economic, and thus were sympathetic to Keynes's policy efforts: but they were suspicious of his political ties to the Liberal Party, and some of the professional; economists were sceptical about his expansionist policies. James Meade and Colin Clark, who were working alongside Keynes, were convinced expansionists by August 1931. Together they were responsible for converting the New Fabians well before 1936. Amongst the sceptics were Gaitskell and Durbin, who were strongly influenced by Hayek's trade cycle theories and who were deeply concerned to demolish 'treasured dogma' within the Labour party, namely the myths that capitalism was collapsing and that socialism could easily replace it and automatically solve the unemployment

problem. As early as 1932 Gaitskell explained why, although 'prosperity' was an important socialist goal, it was not 'the distinguishing characteristic of the Socialist ideal' (Gaitskell 1932).

Meade also played an important role in converting Douglas Jay, whose influential book, *The Socialist Case,* published in 1937, was the first to propose that Keynesian fiscal and monetary measures to control output and employment be explicitly incorporated as part of socialist planning methods. Cole, who thought that the *General Theory* was the most important economics book published since Marx's *Das Kapital* and Ricardo's *Principles,* was quick to point out that because Jay gave such a low priority to nationalization his book contained very little of 'what most people habitually think of as socialism' (Cole 1935). Thus, the introduction of Keynesian methods also served to weaken the case for public ownership as the basis of the socialist economic alternative.

By the late 1930s most democratic socialists in Britain had recognized the importance of the Keynesian message for socialism, and by the end of the war the Labour Party had officially adopted a Keynesian full employment policy. The new macroeconomic analysis provided an obvious answer to the problem of dealing with capitalist collapse. It also reinforced distributive goals, since lower-income families had a higher propensity to consume, and it underscored the importance of central planning to control the economy, since only the government had the power to offset insufficient private spending. So compelling were these arguments that they also converted at least one influential Marxist, John Strachey, to the Fabian cause.

Yet Fabian acceptance of Keynes's economics and of Keynes's basic individualism is often overstated, particularly in the pre-war context. Anthony Wright (in Pimlott 1984) has suggested that the Tawney approach to equality is fundamentally different from the liberal philosophy behind Beveridge's welfare state. A similar contrast can be made between Fabian conceptions about economic planning in the 1930s and Keynesian macroeconomic management. Fabians were explicit about their opposition to the capitalist system, which Keynes wanted to repair, but which they wanted to replace. They were emphatic about the need for major reform of Britain's financial institutions and for substantial growth of the public sector; indeed, they believed that both were essential to implement a successful full employment policy. At least one Fabian economist, Evan Durbin, never accepted *The General Theory* model as the solution to all macroeconomic problems; he believed that it failed to explain the trade cycle, and was therefore unsuitable for the long-term growth problems which the socialist state must solve in order to improve upon capitalism's record.

The principles of market socialism grew out of work initiated by Durbin and Gaitskell, who undertook a systematic reconsideration of the Marshallian microeconomic grounds for intervention and the implications for socialist planning. Together with H.D. Dickinson they demonstrated that the market system by definition could neither price collective goods nor reflect the true social value of externalities, and, therefore, that it could not determine the appropriate allocation of resources for their production. They also incorporated the new economics of imperfect competition associated with Joan Robinson to restate the objections to the existing system, which they termed 'monopoly capitalism'. A planning authority would be able to correct these deficiencies and use the principles of optimal allocation to guide its decisions; in other words, neoclassical criteria should serve as the handmaiden to collective decision-making. In the 1930s and 1940s, many Fabians contributed to the further elaboration of these ideas into a socialist economic system based on free choices in the labour market, consumers' sovereignty through market pricing and marginal cost pricing in nationalized industries. The importance of this analysis was that it added strong theoretical arguments for a mixed economy as an explicit complement to the macroeconomic Keynesian ones.

There were, however, other Fabians who found such arguments hard to take and/or to follow. Barbara Wootton, whose planning schemes were an updated version of the Webbian administrative structure, was clear that prices would have to be controlled in the public interest. Even Dalton, who

recognized that planning was not necessarily socialist, still maintained the early Fabian belief that 'Socialism is primarily a question of ownership' (Dalton 1935, p. 247). With more appreciation for the problems of allocative efficiency under socialism, Cole attempted to fashion a different socialist economics, one which was neither Marxist nor neoclassical (Cole 1935). Although his own system remained a rather sketchy attempt to incorporate socialist distributional goals into decisions about production, he had some telling arguments against his neoclassical comrades, pointing out that market prices reflected the existing income distribution, and thus could not provide the proper signals for socialist allocation. His efforts are particularly interesting for the light they throw on the need to mesh social policy with economic planning, and on the problem of applying neoclassical analysis to meet essentially political goals.

By the end of the 1930s, most Fabians had come to accept the necessity for a mixed economy, if only on practical grounds, because the legislation necessary to secure socialism by parliamentary methods could not be accomplished by one Labour government. Government planning was necessary to ensure aggregative and allocative efficiency and to redistribute income and wealth. Control of what were later known as 'the commanding heights' of the economy was essential to implement the planning alternative, and a central authority was required to make sure that sectional interests, such as bankers, business and trade unionists, did not subvert the public good. However, in an important change of emphasis, Durbin and Gaitskell were explicit that their objections to capitalism and to the Marxist alternative were social and political, not economic (Gaitskell 1935; Durbin 1940). The essence of their socialism was social justice as Tawney defined it. In short, the mixed economy was not simply politically expedient, it was central to the economic operation of the democratic socialist state.

The Fabians and the Mixed Economy in Practice

As authors in the *New Fabian Essays* later pointed out, the war substantially altered the balance of power between the government and the private sector. And in comprehensive plans for recovery, the wartime coalition laid the foundations for bipartisan support of full employment, a unified system of social services and educational reform. Thus, when the Labour government took over in 1945, there was not much resistance to its programme or to its Fabian philosophy.

In 1948 the Fabian Society commissioned W. Arthur Lewis to write a pamphlet on 'the economic perplexities of the moment'. These turned out to be so numerous that Lewis ended up writing a short book, *The Principles of Economic Planning* (1949), an influential statement of the revised conception of market socialism. Like Meade in *Planning and the Price Mechanism*, published in the same year, he argued the case for planning on general interventionist grounds, implicitly rejecting the Durbin–Gaitskell notion that only a socialist government could run the economy efficiently, although one might still believe only a socialist government would. To paraphrase Lewis, socialism was not about the state, any more than it was about property; 'socialism is about equality'. There could be many ways to handle property and to plan the economy, which were not inconsistent with socialism (Lewis 1949, pp. 10–11). Lewis argued that the crucial issue was whether the state should operate 'through the price mechanism or in supersession of it'; the real choice was 'between planning by inducement, and planning by direction'. Lewis himself was neutral on the issue, believing that Britain needed some of both. Although insistent that there must be free consumer goods and labour markets, he argued that demand was not sacred and that it should be manipulated in specific markets and in the aggregate to achieve policy goals. Similarly, he did not believe that nationalization should be taken on its merits. Lewis wanted 'more than we have already got' (steel, banking and chemicals were his candidates), but in no circumstances the whole economy; 'a country whose people love freedom will not wish the state to become the sole employer' (p. 104).

Shortly after this book was published in 1949, Cole as chairman of the Society organized a conference to begin to rethink the way forward now

that the main components of the first Fabian stage to socialism were in place. *New Fabian Essays* published in 1952 was the end result of this effort to take account of important societal changes and the Keynesian Revolution. The essayists were all agreed that the British version of the mixed economy was a permanent Fabian accomplishment, and that the Tories would not dismantle the welfare state nor renege on full employment. Yet, despite the enormous gains, substantial inequities remained and new problems emerged: in particular, the great concentrations of bureaucratic power in the public and private sectors which threatened individual freedom. In general terms the way forward was to continue to pursue equality, to improve labour–management relations and to disperse power as much as possible.

However, the Fabians were still united in their dissatisfaction with that system. Although they were clear that the postwar version of welfare capitalism did not meet their conception of socialism, many of the essayists were vague about what they did want. Writing about equality in *New Fabian Essays,* Roy Jenkins explained that a classless society was one 'in which men will be separated from each other less sharply by variations in wealth and origin than by differences in character', but it was impossible to describe 'the exact shape of the goal'. Of contributors to *New Fabian Essays*, only Crosland was willing to be explicit in the negative sense that he specified four policies which would *not* achieve equality; the continued extension of free social services, more nationalization, the proliferation of controls and further redistribution of income by direct taxation. In an important shift, many Fabians had come not only to believe in the mixed economy, but also to accept its current structural form.

Crosland outlined the main features of what he called 'post-capitalist society': he concluded that it was more equal and more planned than before, but that it was still based on unacceptable class divisions. While individual property rights were no longer the essential basis of economic and social power, they still affected the distribution of wealth. He felt that the power of the state had been expanded sufficiently to exert control over the economy: if anything, physical controls should be reduced as they were unpopular and inefficient. Similarly, nationalization had secured government power in the central sectors of the economy, social legislation had ensured a national minimum welfare level, and full employment policies had removed insecurity and demonstrated that central planning could be directed to meet social ends. Keynesian policies were crucial to maintaining this system, but as these were now well understood, Crosland argued that 'the new society may prove to be a very enduring one'. In *The Future of Socialism* (1956), Crosland spelled out his ideas on planning in more detail; he believed its 'essential role' was Keynesian economic management, that the techniques were no longer controversial nor the preserve of any one party, and that political will, not planning theory, were required to plan effectively; 'if socialists want bolder planning, they must choose bolder ministers.'

One lone dissenter from the general Fabian romance with Keynes was G.D.H. Cole. Although enthusiastic about the *General Theory* when it was published, he had become increasingly concerned about these new directions after the war. Indeed, this was precisely why he had initiated the process of rethinking, and why, as the discussions progressed, he resigned his position as chairman of the Fabian Society. In 1950 he published a short book, *Socialist Economics,* which spelled out his disagreements with the new Fabian approach. First, he thought that Keynesian economics was too involved with aggregates and not sufficiently concerned with the structural problems necessary for a socialist economy to replace the capitalist system. As far as he was concerned the new direction provided a diluted form of socialism, which was 'little more than Keynesian Liberalism with frills'. Second, although Cole had advocated using a wide range of industry controls as early as 1929 and was opposed to total public ownership, he was also explicit in rejecting the current version of the mixed economy 'as a permanent resting place'.

See Also

► Social Democracy

Bibliography

Clarke, P. 1978. *Liberals and social democrats*. Cambridge: Cambridge University Press.
Cole, G.D.H. 1935. *Principles of economic planning*. London: Macmillan.
Crosland, A. 1956. *The future of socialism*. London: Cape.
Crossman, R.H., ed. 1952. *New Fabian essays*. London: Turnstile Press. 3rd impression, London: Dent, 1970.
Dalton, H. 1920. *Some aspects of the inequality of incomes in modern communities*. London: Routledge & Kegan Paul.
Dalton, H. 1935. *Practical socialism for Britain*. London: George Routledge.
Durbin, E.F.M. 1940. *The politics of democratic socialism*. London: Routledge & Kegan Paul.
Durbin, E. 1985. *New Jerusalems: The labour party and the economics of democratic socialism*. London: Routledge & Kegan Paul.
Gaitskell, H.T.N. 1932. Socialism and wage policy. Fabian Society Papers, Box J24/ 2 in Nuffield College, Oxford.
Gaitskell, H.T.N. 1935. Financial policy in the transition period. In *New trends in socialism*, ed. G.E.G. Catlin. London: Lovat Dickson & Thompson.
Gaitskell, H.T.N. 1955. The ideological development of democratic socialism in Britain 1955. *Socialist International Information* 5: 52–53.
Harris, J. 1972. *Unemployment and politics*. Oxford: Oxford University Press.
Lewis, W.A. 1949. *The principles of economic planning*. London: Dobson.
Pimlott, B., ed. 1984. *Fabian essays in economic thought*. London: Gower Publishing.
Schumpeter, J.A. 1942. *Capitalism, socialism, and democracy*. New York: Harper & Row. Torchbook edition, 1962.
Shaw, G.B. 1884, 1896. *Fabian Tract No. 2.* (1884) *Tract No. 70* (1896). Quoted in Crosland (1956).
Shaw, G.B., ed. 1908. *Fabian essays in socialism*. New York: Doubleday. edn, 1967.
Wallas, G. 1923. Article in *Morning Post,* 1 January. See Clarke (1978) for further details.
Wright, A.W. 1979. *G.D.H. Cole and socialist democracy.* Oxford: Clarendon Press.

Fabricant, Solomon (Born 1906)

G. H. Moore

Fabricant was born in Brooklyn, New York, on 15 August 1906. He began his association with the National Bureau of Economic Research in 1930, serving as director of research from 1953 to 1965 and continuing as a member of the Board. From 1944 to 1973 he was on the economics faculty at New York University. His economic studies range across a wide field, including productivity and economic growth, national income and capital formation, trends in government activity, and economic accounting under conditions of inflation.

Fabricant's initial work on productivity demonstrated that in industries with large productivity gains, the resulting cost and price reductions have usually been sufficient to cause output and employment to rise faster than in other industries – a conclusion at variance with the common contention that technology, which is often a source of rapid productivity growth, deprives workers of jobs. Fabricant's research also clarified the understanding of productivity gains and losses during business cycles, with systematic effects on the movements in costs and profits, which in turn play an important role in generating recessions and recoveries.

In his investigation of trends in government activity (1952), he showed how economic development in the United States during the first half of the 20th century had fostered a rise in the relative importance of government. Thus, for example, urbanization promoted the demand for municipal services, advances in transportation technology led to government building of roads and airfields, and increases in family income supported government activities in education, public health, welfare, and old-age assistance. By carefully assembling the facts on government functions, types of organization, and use of labour and capital, and developing a reasoned account of the factors that led to their growth or decline over the past 50 years, Fabricant cast a bright light over what was to happen over the following 30 years.

Selected Works

(Except as noted, all were published in New York by the National Bureau of Economic Research)
1938. *Capital consumption and adjustment*.
1940. *The output of manufacturing industries, 1899–1937*.

1942. *Employment in manufacturing, 1899–1939: An analysis of its relation to the volume of production*.
1952. *The trend of government activity in the United States since 1900*.
1958. *Investing in economic knowledge*.
1959. *Basic facts on productivity change*. Occasional paper 63.
1959. *The study of economic growth*.
1969. *Primer on productivity*.
1976. (With others). *Economic calculation under inflation*. Indianapolis: Liberty Press.
1984. *Toward a Firmer Basis of Economic Policy: The Founding of the National Bureau of Economic Research*. Cambridge, Mass.: National Bureau of Economic Research

Factor Analysis

Irma Adelman

Factor analysis is a branch of analysis of variance used to investigate the structure of a data set. Consider a data set x_{ij} resulting from the observation of several variables j on several objects i. If the data set arises from a complex multidimensional process about which little is known a priori statistical analysis of the data itself might profitably be used to gain insights into various characteristics of the processes which generated the data set. In particular, statistical techniques can be used to: (1) search for a simpler representation of the underlying processes which generated the data by reducing the dimension of the variable space in which the objects are represented; (2) look for the interactions among the variables by forming linear clusters of variables; and (3) seek characterizations of the clusters of variables which relate them to the underlying processes which generated the data set being analysed. Factor analysis performs all three functions.

A variety of factor analytic methods has been introduced. They differ in estimation procedures (least squares or maximum likelihood); fitting equation (original data matrix, covariance or correlation matrix); scaling assumption (original or normalized data, type of normalization and in whether the scaling is performed prior to the estimation or as part of the estimation procedure); and in the normalization principles applied to the factor matrix. For a discussion of the relationship between them see Kruskal (1978). Following Kruskal, we start from the original data, derive the covariance matrix and then discuss the procedures applied to it. The basic technical references are Hotelling (1933), Bartlett (1938), Lawley (1940), Lawley and Maxwell (1971), Joreskog (1967), and Joreskog and Goldberger (1972).

Let the variables j characterizing the objects i be measured as deviations from their means. Assume further that the data set x_{ij} was generated by an r-dimensional linear process, with r significantly smaller than the original number of variables J. We are then seeking a representation of x of the form

$$x_{ij} = \sum_r a_{ir} b_{rj} + v_{ij} \tag{1}$$

which, in some sense, comes closest to representing the original data set. In (1) the a_{ir} represent the coefficients, known as 'factor scores', which indicate the 'regression coefficients' of the objects upon each of the r clusters of variables; the b_{rj} represent the coefficients of the variables in each of the r clusters, known as 'factor loadings' or 'factor patterns'. The r clusters of variables are known as factors or components, and represent the coordinates of the lower-dimensional space onto which the data matrix is mapped. In matrix notation, we can write (1) as (2)

$$X = AB + \Sigma \tag{2}$$

where A is the matrix of a_{ir}, B is the matrix of b_{rj} and Σ is a diagonal disturbance matrix with typical element σ_j^2.

One can fit (2) directly, by least squares or by maximum likelihood, or one can form the sample covariance matrix $C = X'X/N$, where N is the number of objects, and fit it instead. If one assumes that: (1) the a_{ij} are random, identically distributed, with mean 0, and independent both of

each other and of the disturbances and (2) applies the normalization T that sets

$$\frac{N-1}{N}\left(AT^{-1}\right)'\left(AT^{-1}\right) = I \tag{3}$$

then the expected value of the sample covariance matrix C is

$$E(C) = B'B + \Sigma^2 \tag{4}$$

This equation can be fitted either by least squares (Hotelling 1933; Anderson 1958; Harman 1960; Joreskog and Goldberger 1972) or by maximum likelihood methods (Lawley 1940; Joreskog 1967), to obtain estimates for b_{rj} and σ_j^2. Once these estimates have been obtained, a_{ir} can be estimated by regression methods from eqn (2) keeping B fixed.

In the least squares approach the matrix B is estimated by extracting the successive eigenvectors of

$$(C - \lambda_r I)b_r = 0 \tag{5}$$

where λ_r is the rth characteristic root and b_r is the rth eigenvector. The rth column of B, b_r, represents the makeup of the rth component in terms of the original, observable variables. Goodness of prediction measures analogous to significance intervals can be derived for the estimates of B by using Stone–Geisser or Tukey-jack-knife methods (Wold 1982).

In the maximum likelihood approach, we form the likelihood function,

$$L = \frac{1}{2}(N-1)\,1\mathrm{n}|\mathrm{C}| - \frac{1}{2}(N-1)\sum_{i,j} x_{ji}x_{ij}C^{ij}/N - 1 \tag{6}$$

where $|\mathrm{C}|$ is the determinant of C, and C^{ij} is the ijth element of C^{-1}. To find the maximum likelihood estimators of B and Σ, we differentiate (6) with respect to the elements of B and Σ and set the resulting equations equal to zero. The maximum equations are then solved simultaneously for B and Σ by applying techniques such as Fletcher-Powell (1963) for the simultaneous optimization of nonlinear equation systems. The maximum likelihood approach was first developed by Lawley (1940); practical estimation techniques for it were developed by Joreskog (1967). The use of maximum likelihood has both advantages and disadvantages: it requires stringent assumptions about the distributions of the parameter set B and the disturbances Σ but it also enables one to estimate confidence intervals on the parameters of B and on the goodness of fit (Lawley and Maxwell 1971; and Jennrich and Thayer 1973).

Both the least squares approach and the maximum likelihood approach yield estimates of B which are not unique since a rigid rotation of B yields the same estimating equations. Several approaches have been proposed for deriving unique estimates. These include normalization assumptions on $A'A$ or $B'B$ and rotation assumptions on of interpretability such as the varimax rotation (Kaiser 1958).

The first applications of factor analysis in the social sciences were in psychology, for which the technique was first developed by Spearman (1904), and used to analyse mental abilities (see Bolton et al. 1973 for a survey). In economics, the first application was to demand analysis (Stone 1945). Stone hypothesized that demand for commodities is explained by three types of influences: national income and own and other prices; social influences affecting tastes and market conditions; and forces peculiar to a particular community. He used a three-factor confluence analysis model, similar to factor analysis, to identify the factors affecting consumer demand. A recent study of market demand employing modern factor analysis is Huang et al. (1980). Stone (1947) and Geary (1948) used factor analysis to study interaction patterns among time series. Using time series representing the components of national income and product in the US, Stone showed that 97.5 per cent of their total variance could be represented by three factors. Banks (1954) used factor analysis in agriculture to predict overall agricultural productivity from crop productivity data on a small number of crops.

The most numerous applications of factor analysis to economics have been in economic development (Adelman and Morris 1967; Rayner 1970; Schilderinck 1969). In a series of studies,

Adelman and Morris investigated the interdependence of economic, social and political phenomena in the development process. Their observations were 74 countries; their variables were typologies representing various aspects of economic, social and political structure. Four factors explained most of the covariance: a modernization factor, which includes indicators of economic and social development; a political development factor; a political leadership factor; and a social and political stability factor. They found that the relative importance of these factors in explaining intercountry differences in growth rates changes systematically with country development levels, with social forces declining in importance and political leadership increasing. Other applications have been to the economics of education (Aigner and Goldberger 1977) and to stock market prices (King 1966).

Recent uses of factor analysis have been in the estimation of the parameters of unobservable variables, defined as variables whose measurable quantities differ from their theoretical counterparts and to error-in-variables models. Other recent advances have been in nonlinear factor analysis (McDonald 1967) and in the dynamic analysis of factor structures (Geweke 1977).

See Also

- ▶ Arbitrage Pricing Theory
- ▶ Principal Components
- ▶ Stone, John Richard Nicholas (1913–1991)

References

Adelman, I., and C.T. Morris. 1967. *Society, politics, and economic development: A quantitative approach*. Baltimore: Johns Hopkins Press.

Aigner, D.J., and A.S. Goldberger (eds.). 1977. *Latent variables in socioeconomic models*. Amsterdam: North-Holland.

Anderson, T.W. 1958. *An introduction to multivariate statistical analysis*. New York: Wiley.

Banks, C. 1954. The factorial analysis of crop productivity: A reexamination of professor Kendall's data. *Journal of the Royal Statistical Society, Series B* 16: 100–111.

Bartlett, M.S. 1938. Methods of estimating mental factors. *Nature* 141: 609–610.

Bolton, B., Hinman, S., and Tuft, S. 1973. *Annotated bibliography: Factor analytic studies 1941–1970,* 4 vols. Fayetteville: University of Arkansas, Arkansas Rehabilitation Research and Training Center. (Tuft did not collaborate on vols 3 and 4.)

Fletcher, R., and M.J.D. Powell. 1963. A rapidly convergent descent method for minimization. *Computer Journal* 6: 163–168.

Geary, R.C. 1948. Studies in relation between economic time series. *Journal of the Royal Statistical Society, Series B* 10: 140–158.

Geweke, J. 1977. The dynamic factor analysis of economic time-series models. In *Latent variables in socioeconomic models*, ed. D.J. Aigner and A.S. Goldberger. Amsterdam: North-Holland.

Harman, H.H. 1960. *Modern factor analysis,* 3rd ed, revised. Chicago: University of Chicago Press, 1976.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24: 417–441, 498–520.

Huang, C.-L., Raunika, R., and Fletcher, S.M. 1980. Estimation of demand parameters based on factor analysis. Paper presented at the *American Agricultural Economics Association Meetings* in Urbana, Illinois.

Jennrich, R.I., and D.T. Thayer. 1973. A note on Lawley's formulas for standard errors in maximum likelihood factor analysis. *Psychometrika* 38: 571–580.

Jøreskog, K.G. 1963. *Statistical estimation in factor analysis: A New technique and its foundation*. Stockholm: Almqvist & Wiksell.

Jøreskog, K.G. 1967. Some contributions to maximum likelihood factor analysis. *Psychometrika* 32: 443–482.

Jøreskog, K.G. 1984. *Advances in factor analysis and structural equation models*. Lanham: University Press of America.

Jøreskog, K.G., and A.S. Goldberger. 1972. Factor analysis by generalized least squares. *Psychometrika* 37: 243–260.

Kaiser, H.F. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23: 187–200.

King, B. 1966. Market and industry factors in stock price behavior. *Journal of Business* 39(Supplement): 139–190.

Kruskal, J.B. 1978. Factor analysis: Bilinear methods. In *International encyclopedia of statistics*, 307–330. New York: Macmillan.

Lawley, D.N. 1940. The estimation of factor loadings by the method of maximum likelihood. *Royal Society of Edinburgh, Section A, Proceedings* 60: 64–82.

Lawley, D.N., and A.E. Maxwell. 1963. *Factor analysis as a statistical method*, 2nd ed. London: Butterworth, 1971.

McDonald, R.P. 1967. Factor interaction in nonlinear factor analysis. *British Journal of Mathematical and Statistical Psychology* 20: 205–215.

Rayner, A.C. 1970. The use of multivariate analysis in development theory: A critique of the approach used by Adelman and Morris. *Quarterly Journal of Economics* 84: 639–647.

Schilderinck, J.H.F. 1969. *Factor analysis applied to developed and developing countries*. Rotterdam: Rotterdam University Press.

Spearman, C.E. 1904. 'General intelligence' objectively determined and measured. *American Journal of Psychology* 15: 201–293.
Stone, R. 1945. The analysis of market demand. *Journal of the Royal Statistical Society, Series A* 108: 286–382.
Stone, R. 1947. On the interdependence of blocks of transactions. *Journal of Royal Statistical Society, Series B* 9: 1–45.
Thurstone, L.L. 1935. *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. Chicago: University of Chicago Press.
Wold, H. 1982. Soft modeling and some extensions. In *Systems under indirect observation*, vol. II, ed. K.G. Jøreskog and H. Wold, 1–54. Amsterdam: North-Holland.

Factor Content of Trade

Donald R. Davis

Abstract

Trade in goods is also implicitly trade in the services of the factors used to produce those goods. This insight underlies the Heckscher–Ohlin–Vanek model of factor service trade, and provides a laboratory to test our theories concerning world general equilibrium. In recent years this theory has undergone close empirical scrutiny. Early tests strongly rejected the simplest variants of the theory. More recent tests have imposed a modest number of additional restrictions suggested by the data. These involve cross-country heterogeneity in productivity, factor prices, consumption patterns, and the incorporation of non-traded goods. With these restrictions, the model fares well.

Keywords

Factor content of trade; Factor price equalization; Heckscher–Ohlin trade theory; Heckscher–Ohlin–Vanek factor content of trade theory; Integrated equilibrium; Total factor productivity

JEL Classifications

B2

International trade is the cross-border exchange of goods (and services), both final and intermediate. These goods are produced with factors of production located in specific countries, hence the trade in these goods is implicitly also trade in the services of the factors used to produce them. This converts the standard Heckscher–Ohlin model of trade in goods into the Heckscher–Ohlin–Vanek (HOV) model of the factor content of trade.

The most important reason to study the factor content of trade is that it provides a laboratory to test our understanding of world general equilibrium. Countries have specific endowments, technologies, tastes, locations, and distributions of incomes (among other characteristics). The simplest statement of general equilibrium is that these elements are supposed to 'hang together' in a coherent way. Tests of the factor content of trade thus become a first test of the adequacy of our understanding of this world general equilibrium. If we should fail to correctly predict the factor content of trade, then we know that our theory seriously misunderstands at least one element of the underlying reality. If our theory does a good job of making sense of the factor content of trade, then this is a suggestion that the main thrust of our theory is working well. This would give us more confidence, then, in using the theory in policy applications.

The canonical Heckscher–Ohlin–Vanek model of factor service trade can be described simply (see Vanek 1968). Assume that there are G goods, each produced under perfect competition with constant returns to scale. Assume as well that there are F primary factors of production with factor markets competitive. Let $\mathbf{A}$ be an input–output matrix that links net output $\mathbf{Y}$ to gross output $\mathbf{X}$ via $\mathbf{Y} = (\mathbf{I} - \mathbf{A})\mathbf{X}$. Let $\mathbf{B}$ be a matrix of direct factor inputs, with dimension $F \times G$, where columns denote factor inputs required to produce a unit output of a single good and rows show factor inputs for a single factor across all goods. Let $\overline{\mathbf{B}} \equiv \mathbf{B}(\mathbf{I} - \mathbf{A})^{-1}$ be the corresponding matrix of direct plus indirect factor inputs, where both primary and intermediate usage represent cost-minimizing choices. Let c be an index for countries, and W represent the aggregate for the world as a whole.

Let technologies for all goods and quality of all factors be common for all countries of the world, and let there be at least as many goods as factors, so that $G \geq$ F. Assume that trade between countries is free, so that goods prices are equalized. Assume that the distribution of world endowments among countries satisfies the requirements to replicate what has been termed the 'integrated equilibrium' (see Helpman and Krugman 1985). In such a case, the division of world endowments between the countries is of no economic consequence, since outputs adjust across countries so that the countries jointly produce exactly the same output and use the same input ratios as they would if the factors were perfectly mobile across countries. Then factor prices will be equalized (FPE), and for all countries $c \in C$, there are common technology matrices: $\mathbf{B} = \mathbf{B}^c$, and $\overline{\mathbf{B}}=\overline{\mathbf{B}}^c$. For country c with gross output vector $\mathbf{X}^c$ and primary input vector $\mathbf{V}^c$, $\mathbf{B}\mathbf{X}^c = \mathbf{V}^c$. We further assume that demand is identical across countries and homothetic. Let $\mathbf{D}^c$ be country c's vector of final goods demand, $\mathbf{Y}^W$ be the world net output vector, and s^c be country c's share of world spending. Then, with free trade equalizing goods prices, $\mathbf{D}^c = s^c\mathbf{Y}^W$. This identifies the demand for goods, and, by pre-multiplying by the common technology matrix $\overline{\mathbf{B}}$, we can convert this to a statement about the factor content of consumption. $\overline{\mathbf{B}}\mathbf{D}^c = s^c\mathbf{V}^W$. Net trade is $\mathbf{T}^c = \mathbf{Y}^c - \mathbf{D}^c$. Hence the prediction of the net factor content of trade is:

$$(\text{HOV})\quad \overline{\mathbf{B}}\mathbf{T}^c = \mathbf{V}^c - s^c\mathbf{V}^W.$$

Early empirical work, such as Bowen et al. (1987) and Trefler (1995), examined this under the assumption that all countries use the technology matrices of the United States. Without reservation, the conclusion of these papers was that the simplest version of the model is an utter failure. Trefler characterized the central failing as the 'mystery of the missing trade'. If we term $\overline{\mathbf{B}}\mathbf{T}^c$ the *measured* factor content of trade and $\mathbf{V}^c - s^c\mathbf{V}^W$ the *predicted* factor content of trade, then the mystery is that the measured factor content of trade is much smaller than that predicted. Much of the subsequent literature has focused on identifying reasons for the mystery of the missing trade and finding solutions for it.

Virtually every assumption underlying the Heckscher–Ohlin–Vanek model is in principle open to question. The strategic issue has been to bring more data to bear on the question in order to identify which of the assumptions is violated most seriously and what amendments to the theory and data work are needed to fit the pieces of the puzzle together.

Various approaches have been considered. Trefler (1993) develops a model that assumes net factor trade is correctly measured, and calibrates factor quality differences across countries that would rationalize the measured trade. This can be thought of as a model of adjusted factor price equalization. While the theoretical model of quality-adjusted factor service trade is an important addition to the toolkit of researchers, this proposed resolution has not fared well empirically (Davis and Weinstein 2003).

Increasingly, researchers moved to a wider set of departures from the standard Heckscher–-Ohlin–Vanek model. These include differences across countries in total factor productivity (TFP); a breakdown in factor price equalization, even adjusted for the TFP differences; specialization in different traded goods within industries; differences in factor input ratios in both traded and non-traded sectors; and costly trade.

Davis et al. (1997) examined the adequacy of assuming a common technology matrix (in this case, that of Japan) for a set of OECD countries. Instead of looking directly at the factor content of trade, $\overline{\mathbf{B}}^{Japin}\mathbf{T}^c = \mathbf{V}^c - s^c\mathbf{V}^W$, they looked at the factor content of production for these countries, that is, $\mathbf{B}^{Japin}\mathbf{X}^c = \mathbf{V}^c$. This is such a poor fit in the data that they conclude that much of the problem lies in cross-country differences in technology matrices. They went on to develop a theory to predict the factor content of Japanese regions under the assumption that these share FPE, even though Japan does not share FPE with the world as a whole. In this sample, this largely eliminated the mystery of the missing trade.

This left open the larger question of why technologies differed, how they differed, and whether a parsimonious set of departures from the HOV

theory could get the model of factor service trade to work well. Davis and Weinstein (2001a) brought a great deal more data to bear on the problem, developing technology matrices for ten rich OECD countries and a composite rest of the world. Technologies differed systematically, even among these rich OECD countries, so that more capital-abundant countries use more capital-intensive methods industry by industry. As it turned out, this happened in both traded and non-traded goods sectors, the latter being important in identifying a breakdown in relative FPE (because there is less likelihood of aggregation issues impinging). Moreover, recognizing that non-traded sectors in different countries use systematically different input coefficients has a large impact on predicted factor contents. For example, a capital-abundant country uses more capital per worker than would be suggested in an FPE model. For this reason, and because non-traded sectors are large, the capital-abundant country has less 'excess' capital to export through factor services. All told, the adjustments made allow the measured factor content of trade to be approximately 60–80 per cent of that predicted.

The subsequent literature has focused on a number of elaborations and challenges to this work. Feenstra and Hanson (2000) explore in more detail issues of aggregation bias in measurements of net factor service trade. In related work, Davis and Weinstein (2001b) have developed a more elaborate model of gross trade in factor services that helps to understand even North–North trade. In effect, they argue that much of the mystery of the missing trade arose because the focus on net goods trade ignored the fact that when factor intensities are not identical even intra-industry goods trade conveys net factor content. Choi and Krishna (2004) implement alternative tests, based on Helpman (1984), of the net factor content of trade which has the advantage of being robust to breakdowns in FPE, but the disadvantage of needing to have confidence that we can adequately measure the differences in factor prices, including returns to capital. For the sample of bilateral predictions they consider, the model performs well. Reimer (2006) has aimed to incorporate a more elaborate model of trade in intermediates and argues that this diminishes when measured against predicted factor contents.

Research on the factor content of trade is important because it represents the greatest effort on the part of trade economists to assemble all of the pieces of general equilibrium into a single coherent framework relating underlying endowments, production, technology, consumption and trade. The early theoretical and empirical work provided a starting place and a number of anomalies, such as the mystery of the missing trade, that motivated ongoing research. Subsequent literature has gone a long way towards resolving the mystery of the missing trade. But new questions continue to arise, particularly related to trade in intermediates and issues of aggregation. No doubt these will invite further investigation.

See Also

- ▶ Comparative Advantage
- ▶ Factor Price Equalization (Historical Trends)
- ▶ Factor Prices in General Equilibrium
- ▶ Heckscher–Ohlin Trade Theory
- ▶ International Economics, History of
- ▶ Tradable and Non-tradable Commodities

Bibliography

Bowen, H.P., E.E. Leamer, and S. Leo. 1987. Multicountry, multifactor tests of the factor abundance theory. *American Economic Review* 77: 791–809.

Choi, Y.S., and P. Krishna. 2004. The factor content of bilateral trade: An empirical test. *Journal of Political Economy* 112: 887–914.

Davis, D.R., and D.E. Weinstein. 2001a. An account of global factor trade. *American Economic Review* 91: 1423–1454.

Davis, D.R., and D.E. Weinstein. 2001b. Do factor endowments matter for North–North trade? Working Paper No. 8516, NBER, Cambridge, MA.

Davis, D.R., and D.E. Weinstein. 2003. The factor content of trade. In *Handbook of international trade*, ed. J.C. Harrigan. London: Blackwell.

Davis, D.R., D.E. Weinstein, S. Bradford, and S. Kazushige. 1997. Using international and Japanese regional data to determine when the factor abundance theory of trade works. *American Economic Review* 87: 421–446.

Feenstra, R.C., and G.H. Hanson. 2000. Aggregation bias in the factor content of trade: Evidence from U.S. manufacturing. *American Economic Review* 90: 155–160.

Helpman, E. 1984. The factor content of foreign trade. *Economic Journal* 94: 84–94.

Helpman, E., and P. Krugman. 1985. *Market structure and foreign trade*. Cambridge, MA: MIT Press.
Reimer, J.J. 2006. Global production sharing and trade in the services of factors. *Journal of International Economics* 68: 384–408.
Trefler, D. 1993. International factor price differences: Leontief was right! *Journal of Political Economy* 101: 961–987.
Trefler, D. 1995. The case of the missing trade and other mysteries. *American Economic Review* 85: 1029–1046.
Vanek, J. 1968. The factor proportions theory: The N-factor case. *Kyklos* 21: 749–756.

Factor Misallocation and Development

Diego Restuccia

Abstract

The large differences in income per capita across countries are mostly explained by differences in total factor productivity (TFP). This article summarises the evidence on the importance of resource allocation across productive units in explaining the observed differences in TFP across countries.

Keywords

Distortions; Heterogeneous establishments; Misallocation; Productivity

JEL Classification

D4; D10; O1

Introduction

A fundamental question in growth and development economics is why some countries are rich and others poor. To illustrate the enormous differences in income per capita across countries, consider that the average gross domestic product (GDP) per capita of the richest 10% of countries in 2000 was a factor of 40 higher than that of the poorest 10% of countries. In other words, the average person in a rich country produces in just over 9 days what the average person in a poor country produces in an entire year. What are the factors that can explain this enormous difference in standard of living across the world today? Considerable progress has been made in diagnosing the proximate sources of the variation in income per capita across countries with differences in total factor productivity (TFP) considered the dominant factor (see for instance Klenow and Rodriguez-Clare (1997), Prescott (1998) and Hall and Jones (1999)).

The key question is then: what are the sources of low TFP in poor countries? The literature has emphasised the possibility that resources may not be efficiently distributed across production opportunities, thereby generating lower TFP. Such a perspective has received substantial attention in the literature, in terms of both empirical and quantitative work. This perspective has tremendous appeal in understanding productivity differences across countries for at least two reasons. First, in rich economies it is well established that the reallocation of factors across productive units explains a large portion of productivity growth over time. For example, Baily et al. (1992) show that 50% of the growth in manufacturing productivity in the USA in the 1970s and 1980s is attributable to the reallocation of factors across plants, from contracting less-productive plants to expanding moreproductive plants, and from failing plants that exit to entering new plants (see also Foster et al. 2008). Second, it is widely recognised that a number of policies and institutions prevalent in poor countries can distort the allocation of factors across productive units. This is what the literature broadly refers to as misallocation. For instance, it is emphasised that credit markets in poor countries do not operate as efficiently as in rich countries (credit market institutions) and that imperfections in credit markets act as a barrier to the efficient allocation of resources across production opportunities. Similarly, imperfections in land market institutions and labour market institutions can create misallocation. It is also recognised that certain policies (whether intentional or not) can create misallocation as they often effectively apply differently to heterogeneous producers.

The fact that we can produce a long list of factors that can cause misallocation does not immediately imply that misallocation is quantitatively important in explaining low TFP in poor countries. The literature has made substantial progress in empirically documenting the extent of misallocation in poor countries as well as assessing its productivity implications. In addition, the literature has explored many specific factors generating misallocation as well as mechanisms that can amplify their effects on aggregate productivity. In this article, I attempt to synthesise this literature by first describing a very simple model of misallocation. I then follow Restuccia and Rogerson (2013) in classifying the literature into two broad categories. The first is the indirect approach, which provides broad evidence of misallocation and a quantitative assessment of their effect on aggregate TFP. This approach is often silent about the underlying channels through which misallocation takes place. The second is the direct approach, which consists of analysing a particular policy/ institution and making a quantitative assessment of its importance in generating misallocation and low TFP.

A Simple Model of Misallocation

Consider the following simple static economy with production heterogeneity in the spirit of Lucas (1978) and Hopenhayn (1992). A single good is produced. The production unit is an establishment, indexed by i, that produces output according to $y_i = z_i n_i^{\gamma}$, where z_i is establishment-level total factor productivity, n_i is the labour input chosen by the establishment, y_i is the amount of output produced and $\gamma \in (0,1)$. While in practice establishments may differ in many dimensions, I will focus on exogenous differences in z_i. There is a large number of establishments and a measure one of homogeneous workers that supply labour inelastically to the market. For simplicity, assume that there is a finite number of potential z_is. Establishments operate in competitive labour and output markets. Let the price of output be normalised to one and denote the wage rate by w. Given prices, an establishment maximises profits by choosing the labour input. That is,

$$\pi_i(z_i) = \max_{n_i} \{y_i - w n_i\}.$$

The first-order condition for profit maximisation from this problem is given by

$$\gamma z_i n_i^{\gamma-1} = w, \tag{1}$$

which implies that the optimal demand for labour given w is

$$\overline{n}_i(z_i) = \left(\frac{z_i \gamma}{w}\right)^{1/(1-\gamma)}. \tag{2}$$

Note that with all establishments facing the same technological parameters (in this simple case, γ) and prices (w), the more productive establishments (higher z_i) are larger; that is, demand more labour, produce more output and generate more profits. In fact, note from Eq. (2) that the ratio of employment between two establishments i and j is a monotonic function of the ratio of their idiosyncratic productivity $n_i/n_j = \left(z_i/z_j\right)^{1/(1-\gamma)}$. In this setup, establishments have an optimal size which is determined by their idiosyncratic productivity and aggregate factors such as the wage rate. Total output in this economy is the aggregate of output from individual establishments. TFP is the ratio of total output to total labour input. Since total labour is normalised to 1, total output and TFP are the same in this economy. It is easy to show that, in this environment, the allocation from the competitive equilibrium (which in addition includes a wage rate that clears the labour market $\sum_i \overline{n}_i(z_i) = 1$) coincides with the efficient allocation.

I now introduce distortions into this economy in the spirit of Restuccia and Rogerson (2008). While in principle there are many policies/institutions that can create misallocation, it is convenient for the purpose of illustration to generate misallocation via tax/subsidy schemes. Consider then the situation where establishments face a tax/subsidy to output τ_i, where $\tau_i > 0$ means a tax and $\tau_i > 0$ a subsidy. Importantly, establishments will face different τs. I will refer to these policies as idiosyncratic distortions, as in Restuccia and Rogerson (2008), to emphasise the fact that it is precisely the differential tax

F

rates that will create misallocation in this economy. Without entering into the discussion of how the taxes are related to productivity, note that the problem of the establishment now renders a first-order condition which is given by

$$(1-\tau_i)\gamma z_i n_i^{\gamma-1} = w, \tag{3}$$

which implies a demand for labour,

$$\overline{n}_i(z_i,\tau_i) = \left(\frac{(1-\tau_i)z_i\gamma}{w}\right)^{1/(1-\gamma)}. \tag{4}$$

Hence, conditional on productivity, establishments that are taxed more heavily are smaller than establishments that are taxed less. Whereas in the undistorted economy all establishments with the same productivity are of the same size, in the distorted economy some establishments are larger than others on the basis of the distortions alone and that entails an inefficiency. More importantly, whereas in the undistorted economy more productive establishments are larger and as a result have a larger fraction of labour and output, in the distorted economy that is not necessarily the case. Note that from Eq. (4) the ratio of employment between two establishments now depends also on the tax rates faced by these establishments. An unproductive establishment (low z_i) can be large (high n_i) if its τ_i is sufficiently low. Similarly, a productive establishment (high z_i) can be small if its τ_i is sufficiently high. Incidentally, for this reason it is misleading to look only at the size distribution of establishments across countries to make inferences about the differences in the distribution of establishment-level productivity across countries.

Restuccia and Rogerson (2008) emphasise that, given a policy distortion characterised by the function $P(\tau_i, z_i)$ whereby tax/subsidies may be related to establishment productivity, if the policy is such that taxes are applied more heavily to the higher-productivity producers, then the productivity loss associated with that policy will be larger. Much of the direct approach that I will describe later is about measuring and assessing quantitatively policies of this sort.

Up to this point (and in much of the existing literature), misallocation is a narrow, static concept that refers to the reallocation of a given set of aggregate factors across a fixed set of heterogenous productive units. However, I emphasise that, broadly understood, misallocation can also generate negative effects on aggregate factors (for instance on the accumulation of physical and human capital) as well as on the distribution of establishment-level productivity in the economy itself. I will discuss these broader implications of misallocation later. While in this article I emphasise factor misallocation across micro-economic units within a sector, other forms of misallocation can also play a role, such as factor misallocation across sectors, across geographical areas, and across government versus privately owned enterprises (see for instance Restuccia et al. 2008; Restuccia 2011; Brandt et al. 2013).

The Indirect Approach

The indirect approach aims at measuring the full extent of misallocation in an economy without detail as to what policies or institutions may be causing it. Hsieh and Klenow (2009) is a seminal contribution providing empirical measures of misallocation. To illustrate their empirical strategy in the simple framework just discussed, note that in an undistorted economy the marginal product of labour is equalised across all establishments. That is, more productive establishments hire more labour precisely to reduce the marginal product of labour down to the given wage rate (see Eq. 1). In a distorted economy, the marginal product of labour is not equal across establishments that face idiosyncratic distortions. That is, in the distorted economy establishments equate the marginal product of labour to the taxadjusted wage rate, which would not be equal across establishments. While their empirical exercise is obviously more involved than this, in a nutshell, given micro data on productivity z_i and employment ni for individual establishments, we can use Eq. (1) to assess the extent to which the marginal product of labour does not equalise across establishments. To put it

differently, we can use Eq. (3) to calculate the wedges required (the τs) for optimisation to hold. Hsieh and Klenow (2009) use data for China, India and the USA and find large deviations in marginal products, with much larger and systematic differences across establishments in India and China than in the USA. What are the productivity implications of the larger wedges in China and India relative to the United States? Using the model, we can evaluate the quantitative impact of those deviations. It can be shown in the simple framework that whereas efficient allocation results in aggregate TFP as a geometric average of establishment productivity, in the distorted economy, aggregate TFP is lowered by the distortions. Hsieh and Klenow (2009) derive a similar relationship in their more elaborate model, which includes capital, differentiated products and industries, and show that the TFP gains from moving to the efficient allocation of factors are very large in both India and China and much larger than in the USA. More specifically, their results show that by reducing the wedges in India and China to those of the USA, manufacturing TFP in China and India could increase by 30–60%.

A perhaps expected but nevertheless interesting by-product of the micro data is the implied distribution of establishment-level productivity in China, India and the USA. The data show that the distributions of establishments in China and India contain more establishments with lower productivity compared to the distribution in the USA. The data also show that the distributions in China and India contain mass of establishments at extremely low levels of productivity, levels for which there is no mass of establishments in the US distribution. Whereas misallocation focuses on the allocation of factors given the distribution of productivities in a country, an ambitious and very important aspect of the literature is to understand the differences in the distribution of establishment-level productivity and their potential connection to misallocation. I will return to this issue below.

The results from Hsieh and Klenow (2009) have influenced a large body of subsequent work applying similar strategies in a variety of different contexts and country experiences. Broadly speaking, the subsequent literature has confirmed the importance of misallocation in understanding productivity differences – see for instance the work of Busso et al. (2013) for Latin American countries as well as Kalemli-Ozcan and Sorensen (2012) for countries in Africa (see also a more complete review in Restuccia and Rogerson (2013)).

Following an alternative strategy, Bartelsman et al. (2013) provide additional empirical evidence of misallocation and a quantitative assessment for a set of OECD countries. These authors emphasise the covariance between firm-level productivity and firm size as a critical statistic of misallocation. For instance, note that in the simple framework of the previous section, the covariance between establishment productivity and establishment size is high in the undistorted economy, whereas this covariance is diminished in the distorted economy. Their results confirm the important role that misallocation plays in understanding aggregate productivity differences across OECD countries.

The Direct Approach

The direct approach aims to identify specific policies and institutions that generate idiosyncratic effects and misallocation. What policies and institutions are important in generating idiosyncratic effects and misallocation? As alluded to earlier, there is a long list of potential policies and institutions that can create misallocation and reduce aggregate TFP. But the key question is which of these policies and institutions are most responsible for low TFP in poor countries. The approach in the literature has been to select a particular policy or institution that can be measured in the data and to use a model to assess its quantitative effect on productivity. By narrowing the extent of misallocation to a single policy, the studies following the direct approach find much smaller productivity effects than the indirect approach, with productivity losses typically in the range of 5–30%. One important exception is the work of Adamopoulos and Restuccia (2014) where direct empirical measures of idiosyncratic price

distortions in the agricultural sector generate much larger productivity losses (differences in productivity of more than tenfold).

Although with a different emphasis, Hopenhayn and Rogerson (1993) is an early example of this direct approach, where firing taxes are shown to reduce aggregate productivity when establishment productivity varies over time. Firing taxes are a good example of a policy or labour market institution that can create idiosyncratic effects even though the policy is meant to be applied to all establishments lowering their employment level. To see this, note that the firing tax creates a wedge in the downward adjustment of employment (establishments do not lay off as many workers as they would without the tax) as well as a wedge in the upper adjustment (a high level of productivity does not command an increase in employment as large as it would without the tax because of expected mean reversion of the shock). Moreover, in many contexts, such as those of many European countries, firing taxes are applied only to firms with more than a certain number of workers. Since larger firms are associated with higher productivity in an undistorted setting, this exemption of small firms from firing taxes amounts to an idiosyncratic distortion where more productive firms are taxed more heavily than low productivity firms, generating a redistribution of factors from more to less productive establishments and lowering aggregate productivity.

Size-dependent policies – policies that explicitly or implicitly treat producers differently based on the size of the establishment – abound, and Guner et al. (2008) provide both a documentation of these policies as well as a quantitative assessment of how damaging they are for productivity.

Other institutional features, such as the functioning of credit markets or enforcement, can also create idiosyncratic effects. For instance, Banerjee and Duflo (2005) emphasise the role of credit constraints in generating a wide dispersion in the marginal product of capital across firms in India as a likely explanation for low aggregate TFP in that country. Buera et al. (2011) and Greenwood et al. (2013) show how cross-country differences in credit market imperfections distort the allocation of factors to generate large productivity losses. Cross-country differences in property rights can create idiosyncratic effects, as in Ranasinghe (2012). Sometimes even policies that are not intended to have an idiosyncratic impact in effect do, such as trade policies and regulations. For instance, Bond et al. (2013) document the idiosyncratic effects created by the passage of the Smoot-Hawley Tariff Bill during the Great Depression in the USA, while Eslava et al. (2013) study the selection effects in aggregate productivity of a trade reform in Colombia. Leal (forthcoming) studies the effects of the myriad of regulations that determine the large size of the informal sector in Mexico. Another important example of policies/institutions generating idiosyncratic effects and misallocation is in the agricultural sector in poor countries. Adamopoulos and Restuccia (2014) study the role of misallocation in agriculture in explaining the small scale of operation in that sector in poor countries and their low productivity. Policies such as progressive taxes and subsidies that favour small-scale production, land market institutions such as inheritance norms, land fragmentation, and land reform, are shown to substantially lower agricultural productivity.

Amplification Mechanisms

In the context of the standard neoclassical model (with a representative firm structure) it is well known that physical and human capital accumulation amplify the effects of differences in TFP on output per capita (see for instance Klenow and Rodriguez-Clare (1997), Manuelli and Seshadri (2006) and Erosa et al. (2010)). Hence capital accumulation amplifies the impact of misallocation on cross-country income differences.

Much less explored is how policies and institutions that create misallocation affect the distribution of establishment productivity, thereby amplifying the effects of misallocation on aggregate productivity. This is a very important

aspect of broadening the potential impact of misallocation. As discussed earlier, the available micro data across a variety of countries show large differences in the productivity distribution of establishments. To illustrate why the differences in establishment-level productivity may be connected to the same policies that create misallocation, notice that if in the simple framework establishments are allowed to invest in their productivity, then the return to this investment is related to the increased value of the establishment with higher productivity. If distortions are such that high-productivity establishments face larger distortions than low-productivity establishments, the policy also creates a disincentive to invest in productivity by lowering the return to productivity investment. This is what Restuccia (2013) and Bello et al. (2011) do in extending the framework of Restuccia and Rogerson (2008) to understand low productivity in Latin American economies, and is the subject of more elaborate analyses in Ranasinghe (2013), Bhattacharya et al. (2013), Gabler and Poschke (2013) and Hsieh and Klenow (2012). Jones (2011) proposes an amplification mechanism for misallocation that is based on the input–output structure of the economy, as the outputs of many firms are used as inputs in other firms.

Conclusions

Income per capita and total factor productivity differ greatly across countries. Understanding the proximate causes of this variation is a challenging goal in the literature of growth and development, with important welfare and policy implications. Much progress has been made in the literature, as briefly summarized in this article, but further exciting work remains to be done.

See Also

- ▶ Development Economics
- ▶ Economic Growth
- ▶ Inequality Between Nations

Acknowledgments I thank Tasso Adamopoulos and Margarida Duarte for helpful comments. All errors and omissions are my own.

Bibliography

Adamopoulos, T., and D. Restuccia. 2014. The size distribution of farms and international productivity differences. *American Economic Review* 104(6): 1667–1697.

Baily, M., C. Hulten, and D. Campbell. 1992. *Productivity dynamics in manufacturing plants*, 187–267. Microeconomics: Brookings Papers on Economic Activity.

Banerjee, A., and E. Duflo. 2005. Growth theory through the lens of development economics. *Handbook of economic growth*, Vol. 1A, Chapter 7. North-Holland.

Bartelsman, E., J. Haltiwanger, and S. Scarpetta. 2013. Cross-country differences in productivity: The role of allocation and selection. *American Economic Review* 103(1): 305–334.

Bello, O., J. Blyde, and D. Restuccia. 2011. Venezuela's growth experience. *Latin American Journal of Economics* 48(2): 199–226.

Bhattacharya, D., N. Guner, and G. Ventura. 2013. Distortions, endogenous managerial skills and productivity differences. *Review of Economic Dynamics* 16(1): 11–25.

Bond, E., M. Crucini, T. Potter, and J. Rodrigue. 2013. Misallocation and productivity effects of the Smoot–Hawley tariff. *Review of Economic Dynamics* 16(1): 120–134.

Brandt, L., T. Tombe, and X. Zhu. 2013. Factor market distortions across time, space and sectors in China. *Review of Economic Dynamics* 16(1): 39–58.

Buera, F., J. Kaboski, and Y. Shin. 2011. Finance and development: A tale of two sectors. *American Economic Review* 101(5): 1964–2002.

Busso, M., L. Madrigal, and C. Pages. 2013. Productivity and resource misallocation in Latin America. *B. E. Journal of Macroeconomics* 13(1): 1–30.

Erosa, A., T. Koreshkova, and D. Restuccia. 2010. How important is human capital? A quantitative theory assessment of world income inequality. *Review of Economic Studies* 77(4): 1421–1449.

Eslava, M., J. Haltiwanger, A. Kugler, and M. Kugler. 2013. Trade and market selection: Evidence from manufacturing plants in Colombia. *Review of Economic Dynamics* 16(1): 135–158.

Foster, L., J. Haltiwanger, and C. Syverson. 2008. Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review* 98(1): 394–425.

Gabler, A., and M. Poschke. 2013. Experimentation by firms, distortions, and aggregate productivity. *Review of Economic Dynamics* 16(1): 26–38.

Greenwood, J., J. Sanchez, and C. Wang. 2013. Quantifying the impact of financial development on economic development. *Review of Economic Dynamics* 16(1): 177–193.

Guner, N., G. Ventura, and Y. Xu. 2008. Macroeconomic implications of sizedependent policies. *Review of Economic Dynamics* 11(4): 721–744.
Hall, R., and C. Jones. 1999. Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* 114(1): 83–116.
Hopenhayn, H. 1992. Entry, exit, and firm dynamics in long run equilibrium. *Econometrica* 60: 1127–1150.
Hopenhayn, H., and R. Rogerson. 1993. Job turnover and policy evaluation: A general equilibrium analysis. *Journal of Political Economy* 101(5): 915–938.
Hsieh, C., and P. Klenow. 2009. Misallocation and manufacturing TFP in China and India. *Quarterly Journal of Economics* 124(4): 1403–1448.
Hsieh, C., and P. Klenow. 2012. The life-cycle of plants in India and Mexico. *Working Paper 18133*. National Bureau of Economic Research.
Jones, C. 2011. Misallocation, economic growth, and input–output economics. *Working Paper 16742*. National Bureau of Economic Research.
Kalemli-Ozcan, S., and B. Sorensen. 2012. Misallocation, property rights, and access to finance: evidence from within and across Africa. *Working Paper 18030*. National Bureau of Economic Research.
Klenow, P., and A. Rodriguez-Clare. 1997. The neoclassical revival in growth economics: Has it gone too far? In *NBER Macroeconomics Annual,* ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.
Leal, J. 2014. Tax collection, the informal sector, and productivity. *Review of Economic Dynamics* 17(2): 262–286.
Lucas, R. 1978. On the size distribution of business firms. *Bell Journal of Economics* 9: 508–523.
Manuelli, R., and A. Seshadri. 2006. *Human capital and the wealth of nations*. Unpublished manuscript.
Prescott, E.C. 1998. Needed: A theory of total factor productivity. *International Economic Review* 39: 525–552.
Ranasinghe, A. 2012. *Property rights, extortion and the misallocation of talent*. Unpublished manuscript, University of Manitoba.
Ranasinghe, A. 2013. *Impact of policy distortions on firm-level innovation, productivity dynamics and TFP.* Unpublished manuscript, University of Manitoba.
Restuccia, D. 2011. Recent developments in economic growth. *Economic Quarterly Federal Reserve Bank of Richmond* 97(3): 329–357.
Restuccia, D. 2013. The Latin American development problem: An interpretation. *Economia* 13(2): 69–100.
Restuccia, D., and R. Rogerson. 2008. Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic Dynamics* 11(4): 707–720.
Restuccia, D., and R. Rogerson. 2013. Misallocation and productivity. *Review of Economic Dynamics* 16(1): 1–10.
Restuccia, D., D. Yang, and X. Zhu. 2008. Agriculture and aggregate productivity: A quantitative cross-country analysis. *Journal of Monetary Economics* 55(2): 234–250.

Factor Models

Jushan Bai

Abstract

Factor models explain correlations among a set of variables. By postulating that the variables are linked with a small number of latent components, factor models imply a particular structure for the correlation matrix. This article discusses the model's identification and estimation as well as their applications in economics.

Keywords

Approximate factor model; Arbitrage pricing theory; Cointegration; Common factors; Covariance matrix; Diffusion index forecasting; Dynamic factors; Factor analysis; Factor models; Forecasting; Principal components analysis; Sample correlation matrix; Unit roots

JEL Classifications

F

The primary objective of factor analysis is to explain, in a parsimonious way, the correlation among a set of variables. For example, cross-sectional correlation of asset returns may be explained by a single factor, according to capital asset pricing theory (Sharpe 1964; Lintner 1965). The correlation among a large number of macroeconomic variables could be explained by some common shocks (Sargent and Sims 1977; Bernanke and Boivin 2003). Historically, factor models were used by psychologists to examine correlations among a set of test scores. Students' performance across different subjects (maths, philosophy, history, and so on) may potentially be accounted for by a single factor (for example, overall intelligence) (see Lawley and Maxwell 1971). In these examples, a common theme is that a large number of variables are linked with a

small number of unobservable variables which give rise to the cross correlations.

While sample correlation matrix may serve the same purpose in describing the linkage of variables, it is neither parsimonious nor reliable when the number of variables is large relative to the number of observations. Suppose there are N variables, each with T observations. The sample correlation matrix estimates $N(N-1)/2$ parameters without any restriction. When the number of variables exceeds the number of observations ($N > T$), the sample correlation matrix is not of full rank, even though the underlying true correlation matrix is positive definite. A factor model with, say, a single factor attempts to explain the correlation with far fewer parameters, and the resulting correlation matrix will be positive definite. If a factor structure truly (or approximately) characterizes the data generating process, the estimated correlation matrix implied by the factor model constitutes a better estimate than the sample correlation matrix. Even if the data generating process does not follow a factor model, under large N, shrinking the sample correlation matrix towards a correlation matrix with a factor structure may be desirable, in light of Ledoit and Wolf (2003). Most importantly, sample correlation is purely statistical, but factor models have structural interpretations.

In this article, we first present the mathematical form of the factor model, then we state the assumptions employed by classical factor analysis, in which the statistical theory is developed under a fixed N. We then go on to discuss modern factor analysis in which both N and T are large, and in particular, the number of variables (N) can be much larger than the number of observations (T). In each case, we discuss issues related to identification and estimation, and the determination of number of factors. More attention is paid to modern factor analysis. We also present a few applications of factor models in economics, including diffusion index forecasting, panel unit root and cointegration analysis. Finally, we briefly highlight the difference between principal component analysis and factor analysis.

The Model

A factor model takes the form

$$X_{it} = \mu_i + \lambda_i' f_t + e_{it}; i = 1, 2, \ldots, N; \\ t = 1, 2, \ldots, T$$

where X_{it} is the observation on variable i at period t; μ_i is the mean of X_{it} , $\lambda_i(r \times 1)$ is vector of factor loadings, $f_t(r \times 1)$ is vector of factor processes, and e_{it} is the idiosyncratic error term. For example, X_{it} may represent the output growth rate for country i in quarter t, μ_i is the mean growth rate, f_t is a vector of common shocks (technology shocks, financial crises, oil price shocks, and so on) that influence output, λ_i represents the impact of shocks on country i, and e_{it} is the country-specific growth rate. As a further example, X_{it} is the return of asset i in period t, μ_i is the mean return, f_t is a vector of factor returns with zero mean (risk premia adjusted), λ_i is a vector of factor loadings, e_{it} is the idiosyncratic return. The arbitrage pricing theory of Ross (1976) implies restrictions between μ_i and λ_i.

Introducing the following notation

$$X_t = \begin{bmatrix} X_{1t} \\ X_{2t} \\ \vdots \\ X_{Nt} \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}, F = \begin{bmatrix} f_1' \\ f_2' \\ \vdots \\ f_T' \end{bmatrix},$$

$$\Lambda = \begin{bmatrix} \lambda_1' \\ \lambda_2' \\ \vdots \\ \lambda_N' \end{bmatrix}, e_t = \begin{bmatrix} e_{1t} \\ e_{2t} \\ \vdots \\ e_{Nt} \end{bmatrix}$$

the factor model can be rewritten as

$$X_t = \mu + \Lambda f_t + e_t; t = 1, 2, \ldots, T. \quad (1)$$

Below we separately discuss classical factor analysis and modern factor analysis; they are based on different assumptions and inferential theory also differs.

F

Classical Factor Analysis

The main assumptions under classical factor analysis are: (i) f_t and e_t are i.i.d. random variables with zero means; (ii) for normalization purposes, $Ef_t f_t' = I_r$, an identity matrix; (iii) e_t and f_t are uncorrelated; (iv) Λ is a matrix of fixed constants.

Let $\Sigma = E(X_t - \mu)(X_t - \mu)'$, the covariance of X_t and let $\Phi = Ee_t e_t'$, the covariance of e_t. It follows that

$$\Sigma = \Lambda\Lambda' + \Phi. \tag{2}$$

Another key assumption under classical factor analysis is that N is fixed. This assumption appears to be at odds with the essence of factor analysis because this analysis was motivated by large N problems. Nevertheless, traditional applications had been on problems of relatively small N. Furthermore, fixed N assumption makes the statistical inference more tractable. For example, Σ is consistently estimable under fixed N (for example, by the sample covariance matrix) as T goes to infinity. Thus for identification purposes, Σ is assumed to be known.

Without further restrictions, the parameter matrices Λ and Φ are not identifiable since Φ alone would have the same number of parameters as the number of equations in (2). Classical factor analysis thus assumes that Φ is a diagonal matrix. This assumption is not too restrictive since correlation among the variables is supposedly explained by the common factors f_t. In addition, a rotational indeterminacy exists for Λ since $\Lambda G (\Lambda G)' = \Lambda\Lambda'$, where G is such that $GG' = I_r$. To remove this rotational indeterminacy, it is often assumed that $\Lambda'\Phi^{-1}\Lambda$ is a diagonal matrix. A diagonal matrix imposes $r(r-1)/2$ number of restrictions. Thus the number of parameters on the right hand side (2) is $N + Nr - r(r-1)/2$. The number of equations in (2) is $N(N+1) = 2$. Thus in order to identify the parameters, we must have

$$\begin{aligned} s &= N(N+1)/2 - N - Nr + r(r-1)/2 \\ &= \left[(N-r)^2 - (N+r)\right]/2 \geq 0. \end{aligned}$$

This is known as the order restriction, meaning that the number of equations must be no smaller than the number of parameters. This implies that for a factor model to be identifiable, N cannot be smaller than three. When N is exactly three, there can only be one factor ($r = 1$). In this case, $s = 0$ and the number of parameters in the factor model coincides with the number of elements in Σ. When this occurs, no simplification is achieved via factor analysis. Nevertheless, structural interpretation of the model is still of interest since it indicates that the three variables are related to a single common component.

Even for $s = 0$, there may not exist solutions for Λ and Φ to satisfy (2) because factor models further restrict non-negativity for the diagonal elements of Φ; see examples in Lawley and Maxwell (1971, pp. 10–11).

For a larger N and small r, we usually have $s > 0$. In this case, overidentification occurs. Model estimation entails finding Λ and Φ to make the distance between S and $\Lambda\Lambda' + \Phi$ small, where S is the sample covariance matrix. The model is usually estimated by the principal-factor analysis or the maximumlikelihood method (see Mardia et al. 1979; Anderson 1984).

A special case is that $\Phi = \sigma^2 I_N$, a scalar multiple of an identity matrix. In this case, the smallest N that permits identification is $N = 2$, with $r = 1$. To see this, consider

$$\Sigma = \lambda\lambda' + \sigma^2 I_2$$

where $\lambda = (\lambda_1, \lambda_2)'$ is a vector. Reparameterize $\lambda\lambda' = \tau^2\delta\delta'$ with $\|\delta\|^2 = \delta'\delta = 1$, where $\tau^2 = \|\lambda\|^2 = \lambda'\lambda$. The two eigenvalues of the matrix on the right hand side are σ^2 and $\sigma^2 + \tau^2$. The eigenvector associated with the larger eigenvalue is simply δ. Thus we can identify σ^2 as the smaller eigenvalue, and τ^2 as the difference between the two eigenvalues. Moreover, δ is the eigenvector associated with the larger eigenvalue of Σ.

On the assumption that the model is identifiable, the estimated factor loadings will be consistent, the limiting distribution can be found in Anderson (1984) under the assumption of fixed N and large T. Given Λ and Φ, the factor scores f_t can also be estimated by either the generalized least squares (GLS) or the Bayesian method. For example, the GLS estimator of f_t is

$\hat{f}_t = \left(\Lambda'\Phi^{-1}\Lambda\right)^{-1}\Lambda'\Phi^{-1}X_t (t = 1, 2, \ldots, T)$. While f_t is unbiased for f_t, it is not consistent since N is fixed, even if Λ and Φ are known. Finally, the number of factors r is determined via hypothesis testing.

Modern Factor Analysis

Modern factor analysis takes model (1) as the starting point, but then proceeds under different assumptions. First, the number of variables N is assumed to be large, and the limit theory is developed under the assumption that both N and T go to infinity. In particular, N can be much larger than the number of observations T. Second, both f_t and e_t can be serially correlated. Third, Φ needs not to be a diagonal matrix, and in fact, none of the off-diagonal elements needs to be zero. Thus the number of parameters in Φ can be as many as the equations. This is called 'approximate factor model' by Chamberlain and Rothschild (1983). The main interest of this large dimensional factor analysis is to estimate r, Λ, and F. One key assumption of the approximate factor model is that the largest eigenvalue of Φ is bounded uniformly in N. This implies that cross-correlations in the idiosyncratic errors must be weak.

Identification and Estimation

Let $X = (X_{it})$ be the $N \times T$ data matrix and $e = (e_{it})$ be the error matrix of the same dimension. Then

$$X = \Lambda F' + e.$$

Here we assume the constant vector μ to be zero, but without assuming F to have zero mean. If $\mu \neq 0$, the demeaned data matrix should be used in the following discussion, and zero mean for F should also be imposed. Now both Λ and F are to be estimated. Since $\Lambda F' = \Lambda A A^{-1} F'$ for an arbitrary invertible matrix $A(r \times r)$. As an arbitrary $r \times r$ invertible matrix has r^2 free parameters, we need to impose r^2 restrictions. We may impose $\frac{1}{T}F'F = \frac{1}{T}\sum_{t=1}^{T} f_t f_t' = I_r$ together with $\Lambda'\Lambda$ being diagonal. Alternatively, we may impose $\frac{1}{N}\Lambda'\Lambda = I_r$ together with $F'F$ being diagonal. Either way, it will uniquely fix Λ and F (up to a column sign change) given the product $\Lambda F'$.

Under the least squares objective function $S(\Lambda, F) = tr[(X - \Lambda F')(X - \Lambda F')']$, the optimal solution $\left(\tilde{F}, \tilde{\Lambda}\right)$ is simply the principal-components estimator. More specifically, under the first set of normalization restrictions, $\tilde{F}$ is the $T \times r$ matrix consisting of the first r eigenvectors (multiplied by $\sqrt{T}$) associated with the first r largest eigenvalues of the $T \times T$ matrix $X'X$, and $\tilde{\Lambda} = \frac{1}{T}X\hat{F}$. Under the second set of identification restrictions, the optimal solution $\left(\overline{F}, \overline{\Lambda}\right)$ is also an eigenvalue problem associated with the matrix of XX', which is $N \times N$. That is, $\overline{\Lambda}$ is the matrix of the first r eigenvectors (multiplied by $\sqrt{N}$) of the matrix XX' and $\overline{F} = X'\overline{\Lambda}/N$ (see Connor and Korajzcyk 1986; Stock and Watson 2002a). The relationships between the two sets of solutions are given by $\tilde{F} = \overline{F}V^{-1/2}$ and $\tilde{\Lambda} = \overline{\Lambda}V^{1/2}$, where V is an $r \times r$ diagonal matrix consisting of the eigenvalues of the matrix $\frac{1}{NT}XX'$. The statistical properties of $\tilde{F}$ and $\tilde{\Lambda}$ are analysed by Bai (2003).

The Number of Factors

The number of factors r can be consistently estimated using the information criterion approach of Bai and Ng (2002). Let $\hat{\sigma}^2(k)$ denote the sum of squares residuals (divided by NT) when k factors are allowed, that is, $\hat{\sigma}(k)^2 = S\left(\tilde{\Lambda}^k, \tilde{F}^k\right)/(TN)$. Consider the following criterion

$$IC(k) = \log\hat{\sigma}(k)^2 + kg(N, T).$$

If $g(N, T)$ is such that $g(N, T) \to 0$ and $\min [N, T]g(N, T) \to \infty$, then $P\left(\hat{k} = r\right) \to 1$, where $\hat{k}$ minimizes the information criterion. For example, $g(N, T) = (N + T) \log (NT)/(NT)$ satisfies the above condition.

Nonstationary Factor Analysis

When the factor process f_t is a vector of I(1) or integrated processes such that $f_t = f_{t-1} + \eta_t$, X_{it} is nonstationary. Examples include nominal exchange rates series (see Banerjee et al. 2005). When the idiosyncratic process e_{it} is I(0) both Λ and F, as well as r can be consistently estimated,

F

as shown by Bai (2004). Since X_{it} (for all i) share the same common stochastic trends f_t, X_{it} are cointegrated among themselves.

When the idiosyncratic process e_{it} is I(1) for all i such that $e_{it} = e_{i,\,t-1} + \varepsilon e_{it}$, there is no cointegration among the observable X_{it}. But still, the common stochastic trends are well defined and can be estimated consistently up to a rotation, a striking contrast with a fixed N spurious system. In a small N system, common stochastic trends and cointegration are synonymous. A spurious system has no common trends or at least cannot be discerned. To see how large N makes a difference, consider the system in differenced form

$$\Delta X_{it} = \lambda_i' \eta_t + \varepsilon_{it}$$

where $\eta_t = \Delta f_t$ and $\varepsilon_{it} = \Delta e_{it}$. This is a standard factor model, and η_t can be estimated under large N and large T. Recumulating η_t will obtain f_t up to a location (unless $f_1 = 0$) and scale shift. When the initial observation $X_{i1} = \lambda_i' f_1 + \varepsilon_{i1}$ is included in estimation, there is only a scale shift in the estimated f_t.

The above idea is implemented in Bai and Ng (2004) for testing panel unit roots. The process X_{it} will have a unit root if either f_t or e_{it} has a unit root. The key is to consistently estimate f_t and e_{it} without knowing a priori their integration orders. Bai and Ng propose to test separately the nonstationarity property for the common component and the idiosyncratic components. This permits us to trace the source of a nonstationary property arising from a common or idiosyncratic component.

Moon and Perron (2004) and Phillips and Sul (2003) propose methods for testing unit roots in the idiosyncratic errors. Related studies can be found in the surveys by Breitung and Pesaran (2008) and Choi (2006).

Diffusion-Index Forecasting

Large-dimensional factor models have proven useful in forecasting macroeconomic variables. Let y_t be the variable to be forecasted, say inflation. Consider the h-period-ahead forecasting equation,

$$y_{t+h} = \alpha' w_t + \beta' f_t + \varepsilon_{t+h}$$

where w_t is a set of observable predictors, such as the lags of y_t and the unemployment rate under the Philips curve model. Here, f_t is not observable, but it captures the co-movement among a large number of macroeconomic variables X_{it}, which links to f_t according to (1). Stock and Watson (2002a, b) suggest that f_t be extracted from X_{it} to obtain $\hat{f}_t$, and then use $\hat{f}_t$ in place of f_t in the forecasting equation. This method is referred to as diffusion index forecasting, which outperforms many competing methods. Bai and Ng (2006a) analyse the statistical properties of this method. A modified diffusion index approach is proposed in Bai and Ng (2006b). The modified approach consists of two steps. The first step selects a subset of X_{it} that is relevant to y_t based on certain criteria. The second step proceeds as the usual diffusion approach using the selected subset of X_{it} only.

Large Dimensional Covariance Matrix

A large dimensional covariance matrix is useful in financial risk management and portfolio construction. For $N > T$, the sample covariance matrix $S(N \times N)$ as an estimator for Σ is not full rank. Thus we consider factor-model based estimator. For this purpose, we use a demeaned data matrix, denoted by X_c (remove the time series mean for each series). Note that the sample covariance matrix is $S = \frac{1}{T-1} X_c X_c'$. Estimate the factor $\tilde{F}$ in the same manner as above using $X_c' X_c$, then $\tilde{\Lambda} = X_c \tilde{F}/T$, and $\tilde{e} = X_c - \tilde{\Lambda}\tilde{F}'$. Given these estimates, a factor-model based covariance matrix is then defined as

$$\hat{\Sigma} = \alpha \hat{\Lambda} \Lambda' + D$$

where D is a diagonal matrix with typical element $\hat{d}_i = \frac{1}{T-1} \sum_{t=1}^{T} \hat{e}_{it}^2 (i = 1, 2, \ldots, N)$, $\alpha = T/(T-1)$. The diagonal elements of $\hat{\Sigma}$ coincide with the corresponding elements of the sample covariance matrix S. In essence, $\hat{\Sigma}$ is an estimator that shrinks the off-diagonal elements of S towards zero. Also the inverse of this matrix is quite easy to compute, it is given by

$$\tilde{\Sigma}^{-1} = D^{-1} - \alpha D^{-1} \tilde{\Lambda} \left(I_r + \alpha \tilde{\Lambda}' D^{-1} \tilde{\Lambda} \right)^{-1} \tilde{\Lambda}' D^{-1},$$

which only requires the inverse of a diagonal matrix and that of an $r \times r$ matrix. Other covariance estimators are discussed by Ledoit and Wolf (2003, 2004), and Fan et al. (2006).

Dynamic Factor Models

In model (1), the factor process f_t is allowed to be a general dynamic process. However, the relationship between X_{it} and f_t is static. A general dynamic factor model is defined as

$$X_{it} = \mu_i + \gamma_i(L)'u_t + e_{it}$$

where u_t are i.i.d. random vectors, and $\gamma_i(L) = \sum_{i,k=0}^{\infty} \gamma_{ik}L^k$ with L being the lag operator. Sargent and Sims (1977), Quah and Sargent (1993) and Geweke and Singleton (1981) are among the early researchers who have studied the dynamic factor models in economics. Identification and estimation of the general dynamic factor model is studied by Forni et al. (2000), who extend the dynamic principal components analysis of Brillinger (1981) to large N. If $\gamma_i(L)$ is a finite order polynomial such that $\gamma_i(L)'u_t = \gamma_{i0}'u_t + \cdots + \gamma_{ip}'u_{t-p}$ (a finite order moving average) then the dynamic factor model can be written as a static factor model by defining $f_t = \left(u_t', \ldots, u_{t-p}'\right)'$, and $\lambda_i = \left(\gamma_{i0}', \ldots, \gamma_{ip}'\right)'$ so that $\gamma_i(L)'u_t = \lambda_i' f_t$. The usual principal components method is still applicable. In general, when the coefficients in $\gamma_i(L)$ decays to zero quickly, $\gamma_i(L)u_t$ can be approximated by a finite order moving average.

Relationship with Principal Components Analysis

Principal components analysis (PCA) seeks linear combinations of the observable variables that give rise to maximum variations. The aim is to summarize the data with as few components as possible without losing too much information. In doing so, it imposes no restrictions on the covariance matrix, as does factor analysis. As such, PCA is a pure dimension-reduction technique.

Factor analysis aims to explain the correlations or co-movements among the observable variables. It assumes that observable variables are linked with a small number of unobservable variables (factors), which are responsible for the correlation. Thus factor analysis is conducted based on a model. In contrast, PCA can be considered as a model-absence method.

Factor models can be estimated by the principal components method. The socalled principal-factor analysis is an iterated principal components method (see Mardia et al. 1979). There are three situations in which the principal components method (without iteration) will give either identical or similar results as other factor estimation methods: (i) the idiosyncratic covariance is a scalar multiple of an identity matrix, that is, $\Phi = \sigma^2 I_N$; (ii) the idiosyncratic error variance is small, that is, Φ is close to zero; (iii) the number of variables of N is large.

Summary

Factor analysis is a model for correlations, postulating that correlations be induced by a few unobservable common components. The model implies a structure on the covariance matrix, which has far fewer free parameters than unrestricted covariance matrix. Therefore, factor models are employed in problems where a reduction in the number of parameters is desired. Applications in economics include modelling cross-sectional correlation, capturing co-movements, forecasting, panel unit root and cointegration analysis, as well as financial risk management and optimal portfolio construction.

See Also

- Arbitrage Pricing Theory
- Cointegration
- Forecasting
- Longitudinal Data Analysis
- Time Series Analysis
- Unit Roots

Bibliography

Anderson, T.W. 1984. *An introduction to multivariate statistical analysis*. New York: Wiley.

Angelini, E., Henry, J. and Mestre, R. 2001. Diffusion index-based inflation forecasts for the Euro area. Working Paper No. 6, European Central Bank.

Bai, J. 2003. Inferential theory for factor models of large dimensions. *Econometrica* 71: 135–173.

Bai, J. 2004. Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics* 122: 137–183.

Bai, J., and S. Ng. 2002. Determining the number of factors in approximate factor models. *Econometrica* 70: 191–221.

Bai, J., and S. Ng. 2004. A panic attack on unit roots and cointegration. *Econometrica* 72: 1127–1177.

Bai, J., and S. Ng. 2006a. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* 74: 1133–1150.

Bai, J. and Ng, S. 2006b. Forecasting using targeted predictors. Unpublished manuscript, Department of Economics, New York University.

Banerjee, A., M. Marcellino, and C. Osbat. 2005. Testing for PPP: Should we use panel methods? *Empirical Economics* 30: 77–91.

Bernanke, B., and J. Boivin. 2003. Monetary policy in a data rich environment. *Journal of Monetary Economics* 50: 525–546.

Breitung, J., and M.H. Pesaran. 2008. Unit roots and cointegration in panels. In *The econometrics of panel data*, 3rd ed., ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

Brillinger, D.R. 1981. *Time series: Data analysis and theory*. San Francisco: Holden-Day.

Chamberlain, G., and M. Rothschild. 1983. Arbitrage, factor structure and meanvariance analysis in large asset markets. *Econometrica* 51: 1305–1324.

Choi, I. 2006. Nonstationary panels. In *The Palgrave handbook of econometrics*, ed. T. Mills and K. Patterson, vol. 1. Basingstoke: Palgrave Macmillan.

Connor, G., and R. Korajzcyk. 1986. Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics* 15: 373–394.

Fan, J., Fan, Y. and Lv, J. 2006. High dimensional covariance matrix estimation using a factor model. Unpublished manuscript, Princeton University.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin. 2000. The generalized dynamic factor model: Identification and estimation. *Review of Economics and Statistics* 82: 540–554.

Geweke, J., and K.J. Singleton. 1981. Maximum likelihood confirmatory factor analysis of economic time series. *International Economic Review* 22: 37–54.

Lawley, D.N., and A.E. Maxwell. 1971. *Factor analysis as a statistical method*. London: Butterworth.

Ledoit, O., and M. Wolf. 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* 10: 603–621.

Ledoit, O., and M. Wolf. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* 88: 365–411.

Lintner, J. 1965. Security prices, risk, and maximal gains from diversification. *Journal of Finance* 20: 587–615.

Mardia, K.V., J.T. Kent, and J.M. Bibby. 1979. *Multivariate analysis*. New York: Academic Press.

Moon, H.R., and B. Perron. 2004. Testing for unit root in panels with dynamic factors. *Journal of Econometrics* 122: 81–126.

Phillips, P.C.B., and D. Sul. 2003. Dynamic panel estimation and homogeneity testing under cross section dependence. *Econometrics Journal* 6: 217–259.

Quah, D., and T. Sargent. 1993. A dynamic index model for large cross sections. In *Business cycles, indicators and forecasting*, ed. J. Stock and M. Watson. Chicago: University of Chicago Press.

Ross, S. 1976. The arbitrage theory of capital asset pricing. *Journal of Finance* 13: 341–360.

Sargent, T., and C. Sims. 1977. Business cycle modelling without pretending to have too much a priori economic theory. In *New methods in business cycle research*, ed. C. Sims. Minneapolis: Federal Reserve Bank of Minneapolis.

Sharpe, W. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.

Stock, J.H., and M.W. Watson. 2002a. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97: 1167–1179.

Stock, J.H., and M.W. Watson. 2002b. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20(2): 147–162.

Factor Price Equalization (Historical Trends)

Kevin H. O'Rourke

Abstract

The Heckscher–Ohlin prediction that international trade should lead to relative factor prices converging internationally is one that receives abundant empirical support for the period that was the focus of interest for these two economists, namely the 'long nineteenth century'. In labour-abundant regions, wage–rental ratios

increased, whereas they declined in land-abundant countries.

Keywords

Factor price equalization; Heckscher, E. F; Heckscher–Ohlin trade theory; Ohlin, B. G; Skill-biased technical change; Transportation costs; Wage–rental ratio

JEL Classification

C30; C31; C32; C33

When twenty-first-century undergraduate economists are taught trade theory, they inevitably encounter the standard Heckscher–Ohlin trade theory, which was for many years, and perhaps still is, the workhorse of international trade theory. The '2 × 2 × 2' version of the theory which they first study surely strikes many of them as unrealistic in the extreme, with its strong predictions of factor price equalization, which logically follows when both countries produce both goods using the same technology. However, the origins of the theory lie in the attempts of two Swedish economists, Eli Heckscher (who was an economic historian) and Bertil Ohlin, to understand the world around them, and in particular to make sense of the global economy of the late nineteenth century. Not surprisingly, perhaps, their theoretical predictions find ample empirical support in the historical records of that time.

Bertil Ohlin presented the theory as follows:

> Australia has a small population and an abundant supply of land, much of it not very fertile. Land is consequently cheap and wages high, in relation to most other countries. It would therefore seem profitable to produce goods requiring large areas of less fertile land but relatively little labour. Such is the case, for example, in wool production . . . Similarly, regions well endowed with technically trained labor and capital will specialize in industrial production . . . Exports from one region to the other will on the whole consist of goods that are intensive in those factors with which this region is abundantly endowed and the prices of which are therefore low . . . In short, commodities that embody large quantities of particularly scarce factors are imported, and commodities intensive in relatively abundant factors are exported. . . . Australia exchanges wool and wheat for industrial products since the former embody much land and little labour while the opposite is true of industrial products. Australian land is thus exchanged for European labor. (Flam and Flanders 1991, p. 90)

He then argued that the level of trade integration helped determine factor prices in both regions:

> If, for example, Australia produced its own industrial products rather than importing them from Europe and America in exchange for agricultural products, then, on the one hand, the demand for labor would be greater and wages consequently higher, and on the other the demand for land, and therefore rent, lower than at present. At the same time, in Europe the scarcity of land would be greater and that of labor less than at present if the countries of Europe were constrained to produce for themselves all their agricultural products instead of importing some of them from abroad. Thus trade increases the price of land in Australia and lowers it in Europe, while tending to keep wages down in Australia and up in Europe. The tendency, in other words, is to approach an equalization of the prices of productive factors. (Flam and Flanders 1991, pp. 91–92)

Three points should be noted about these quotations. First, Ohlin presented the theory using an example that seems to lend itself more easily to formalization via a three-factor, two-good model (in which land, labour and capital produce agricultural and industrial products) than to formalization via the 2 × 2 framework that is often associated with Heckscher–Ohlin theory today. Second, he speaks of a 'tendency' to 'approach an equalization' of factor prices, but not of factor price equality per se. That is, his prediction is that there would be factor price *convergence*. Third, the metaphor that motivated him was one that reflected the international economy of the late nineteenth century, in which intercontinental trade flows for the most part reflected an exchange of resource-intensive products coming from the New World, but also from resource-abundant regions in Asia and Africa, for labour-intensive (and also capital-intensive) manufactured goods produced in western Europe and parts of North America (Findlay and O'Rourke 2007, ch. 7).

The nineteenth century, and particularly the period from roughly 1840 onwards, offers the perfect context in which to study the empirical relevance of such a theory, for it was a period that saw a

F

Factor Price Equalization (Historical Trends), Table 1 Freight factors, 1820–1910 (per cent)

Commodity	From	To	Basis	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910
Wheat	Baltic	UK	Import	8.0	7.1	7.2		6.8	9.6	4.5	3.5	5.9	3.4
Wheat	Black Sea	UK	Import	15.5	16.3			15.0	17.3	9.2	9.7	10.8	6.8
Wheat	East coast, USA	UK	Import		10.3		7.5	10.9	8.1	8.6	5.0	8.2	3.2
Wheat	New York	UK	Export				10.5					6.9	
Wheat	New York	UK	Import				9.4					6.2	
Wheat	Chicago	UK	Export						33.0	21.7	13.3	15.9	7.4
Wheat	South America	UK	Import								15.6	18.5	7.4
Wheat	Rio de la Plata	UK	Import								15.4		6.9
Wheat	Australia	UK	Import								22.3	26.7	15.4
Coal	Britain	Genoa	Export			213.1	224.5	246.1	194.0	163.1	69.7	64.5	53.8
Coal	Nagasaki	Shanghai	Export							84.0	57.0	35.0	20.0
Copper ore	West coast, S. America	UK	Import					21.3			7.8		
Guano	West coast, S. America	UK or European Continent	Import					24.9			18.5		
Nitrate	West coast, S. America	UK or European Continent	Import					34.1			23.0		9.7
Coffee	Brazil	UK or European Continent	Import					5.2			2.0		1.5
Salted hides	Rio de la Plata	UK	Import					3.1			3.8		
Wool	Rio de la Plata	UK	Import					1.3			1.3		

Source: Findlay and O'Rourke (2007, Table 7.2)

dramatic, worldwide decline in transport costs (O'Rourke and Williamson 1999, ch. 3). For example, Knick Harley's (1988) index of British ocean freight rates declines by about 70% between 1840 and 1910, after having remained roughly constant for a century or so. Table 1 presents freight factors (that is to say, transport costs as a percentage of either the import or the export price of a commodity) for several commodities and routes, and the picture which emerges is one of sharply falling transport costs on many routes. The implication is that the relative prices of imported goods should have been steadily declining across continents, as commodity market integration lowered intercontinental price gaps. Furthermore, these declining transport costs were linking continents with very different factor endowments, implying that there should have been scope for trade to have had the sort of impact on factor prices that Heckscher and Ohlin said it should.

In this historical context, one key prediction of the theory is as follows. In labour-abundant regions, such as the crowded countries of Europe, declining transport costs should have led to the relative price of agricultural commodities falling, as they were imported from land-abundant regions, and thus to the ratio of wages to land rents rising. In land-abundant countries, such as the frontier societies of the New World, declining transport costs should have led to the rise of relative price of agricultural commodities, and thus to a fall in wage–rental ratios. Economic historians have examined this prediction at great length. O'Rourke et al. (1996) presented evidence for seven affluent 'Atlantic economy' economies, while more recently Jeffrey Williamson, in a

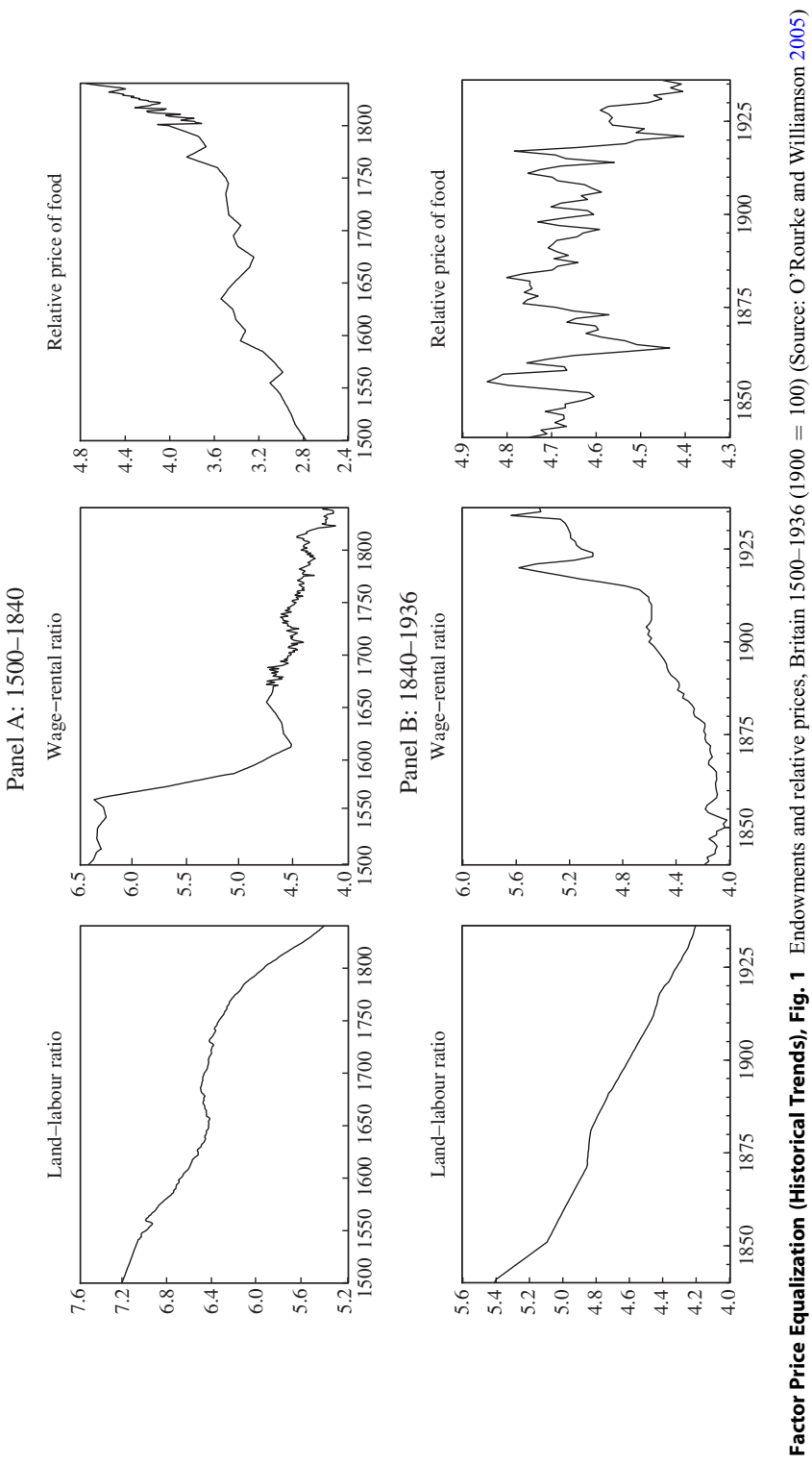
series of papers summarized in Williamson (2002), has expanded the work to include data for several developing economies. By and large, the Heckscher–Ohlin predictions hold good for the late nineteenth century, both for western Europe and the New World, and for those Third World countries that participated in the late nineteenth-century global economy (O'Rourke and Williamson 1999, ch. 4; Williamson 2002, Table 3, p. 73). In land-scarce economies such as those of Japan, Korea, Taiwan or the United Kingdom, the wage–rental ratio increased substantially; while it fell sharply in land-abundant food exporting nations and regions such as Argentina, Uruguay, Burma, Siam, Egypt, the United States, Canada, Australia and the Punjab (see Table 2).

Of course, the fact that wage–rental ratios were systematically trending upwards in labour-abundant economies, and downwards in land-abundant economies, cannot be taken as proof that the Heckscher–Ohlin theory was at work, any more than rising skill premia in many Organisation for Economic Co-operation and Development (OECD) economies can be taken as evidence of factor price equalization today, in 2007. As today's debate about 'trade and wages' suggests, other forces might be at work driving up the ratio of skilled to unskilled wages in skill-abundant economies, most notably perhaps technological change biased in favour of skilled workers. (Recent contributions to the literature on this controversy include Collins 1998; Feenstra 2000; and Feenstra and Hanson 2004.) For the late nineteenth century, both econometrics and simulation exercises indicate that the wage–rental ratio trends documented in Table 2 were indeed in part due to Heckscher–Ohlin forces. That is to say, the price of agricultural products relative to manufactured products was negatively related to the wage–rental ratio during this period (O'Rourke and Williamson 1994; 1999; O'Rourke et al. 1996). It is noticeable that wage–rental ratios increased by less in protectionist economies such as those of France and Germany than in the free-trading United Kingdom. Further evidence in favour of this Heckscher–Ohlin interpretation of nineteenth-century distributional trends comes from a comparison of the pre-1800 and nineteenth-century periods (O'Rourke and Williamson 2005). Before 1800, British land–labour ratios were trending

Factor Price Equalization (Historical Trends), Table 2 Wage–rental ratio trends, 1870–1914 (1911 = 100)

Period	1870–1874	1875–1879	1880–1884	1885–1889	1890–1894	1895–1899	1900–1904	1905–1909	1910–1914
Land-abundant countries or regions									
Argentina			580.4	337.1	364.7	311.1	289.8	135.2	84
Australia	416.2	253	239.1	216.3	136.2	147.7	130	97.9	100.6
Burma					190.9	189.9	186.8	139.4	106.9
Egypt	196.7	174.3	276.6	541.9	407.5	160.1	166.7	64.4	79.8
Punjab		198.5	147.2	150.8	108.7	92	99.8	92.4	80.1
Siam	4699.1	3908.7	3108.1	2331.6	1350.8	301.3	173	57.2	109.8
USA	233.6	195	188.3	182.1	173.5	175	172.4	132.7	101.1
Uruguay	1112.5	891.3	728.3	400.2	377.2	303.6	233	167.8	117.9
Land-scarce countries									
Britain	56.6	61.4	64.9	73.1	79.1	87.3	91.4	98.1	102.7
Denmark	44.8	43.5	44.8	56.6	66.7	87.9	103.8	99.7	100
France	63.5	62.9	67.3	73.8	80.4	91.8	103.2	106.4	99.8
Germany	84.4	80	82.3	86	98	108.2	107.6	104.6	100.2
Ireland	51.3	62.2	72.7	86.4	102.7	122.1	111.2	101.7	94.1
Japan				79.9	68.6	91.3	96.1	110.4	107.5
Korea								102.8	121.9
Spain	42.7	55.8	58.6	73	81.8	85.5	74.9	85.7	86.4
Sweden		43.7	50.7	57.8	65.3	78.6	87.9	92.5	99.1
Taiwan							68.1	85.2	96.6

Source: Williamson (2002, Tables 3 and 4, pp. 73–74)



Factor Price Equalization (Historical Trends), Fig. 1 Endowments and relative prices, Britain 1500–1936 (1900 = 100) (Source: O'Rourke and Williamson 2005)

downwards, as population expanded but land supplies remained relatively constant. In a closed economy setting, this would be expected to lead to an increase in the relative price of agricultural commodities, and to a decline in wage–rental ratios; and this is indeed what happened (Fig. 1, Panel A). From 1840 onwards, by contrast, relative agricultural prices stopped rising, and eventually started falling, while the wage-rental ratio stopped falling and started rising (Panel B). This switch occurred despite an acceleration in British population growth, and is consistent with a British economy opening up to trade and becoming more exposed to the factor price convergence forces identified by Heckscher and Ohlin

Factor prices were certainly not equalized during the late nineteenth century, any more than they have been equalized today. But they converged between continents, in precisely the manner envisaged by Heckscher and Ohlin.

See Also

- ▶ Cliometrics
- ▶ Heckscher–Ohlin Trade Theory
- ▶ International Trade Theory

Bibliography

Collins, S.M. 1998. *Imports, exports, and the American worker*. Washington, DC: Brookings Institution Press.

Feenstra, R.C. 2000. *The impact of international trade on wages*. Chicago: University of Chicago Press and NBER.

Feenstra, R.C., and G.H. Hanson. 2004. Global production sharing and rising inequality: A survey of trade and wages. In *Handbook of international trade*, vol. 1, ed. E. Kwan Choi and J. Harrigan. Oxford: Blackwell.

Findlay, R., and K.H. O'Rourke. 2007. *Power and plenty: Trade, war and the world economy in the second millennium*. Princeton: Princeton University Press.

Flam, H., and M.J. Flanders. 1991. *Heckscher–Ohlin trade theory*. Cambridge, MA: MIT Press.

Harley, C.K. 1988. Ocean freight rates and productivity, 1740–1913: The primacy of mechanical invention reaffirmed. *Journal of Economic History* 48: 851–876.

O'Rourke, K.H., and J.G. Williamson. 1994. Late 19th century Anglo-American factor price convergence: Were Heckscher and Ohlin right? *Journal of Economic History* 54: 892–916.

O'Rourke, K.H., and J.G. Williamson. 1999. *Globalization and history: The evolution of a nineteenth-century atlantic economy*. Cambridge, MA: MIT Press.

O'Rourke, K.H., and J.G. Williamson. 2005. From Malthus to Ohlin: Trade, industrialisation and distribution since 1500. *Journal of Economic Growth* 10: 5–34.

O'Rourke, K.H., A.M. Taylor, and J.G. Williamson. 1996. Factor price convergence in the late nineteenth century. *International Economic Review* 37: 499–530.

Williamson, J.G. 2002. Land, labor and globalization in the third world, 1870–1940. *Journal of Economic History* 62: 55–85.

F

Factor Price Frontier

Heinz D. Kurz

JEL Classifications
D3

The constraint binding changes in the distributive variables, in particular the real wage rate (w) and the rate of profit (r), was discovered (though not consistently demonstrated) by Ricardo: 'The greater the portion of the result of labour that is given to the labourer, the smaller must be the rate of profits, and vice versa' (Ricardo 1971, p. 194). He was thus able to dispel the idea, generated by Adam Smith's notion of price as a sum of wages and profits, that the wage and the rate of profit are determined *independently* of each other. Ever since the inverse relationship between the distributive variables played an important role in long-period analysis of both classical and neoclassical descent. In more recent times it was referred to by Samuelson (1957), who later dubbed it 'factor price frontier' (cf. Samuelson 1962). Hicks (1965, p. 140, n.1) objected that this term is unfortunate, since it is the earnings (quasi-rents) of the (proprietors of) capital goods rather than the rate of profit which is to be considered the 'factor price' of capital (services). A comprehensive treatment of the problem under consideration within a classical framework of the analysis, including joint production proper, fixed capital and scarce natural resources, such as land, was

provided by Sraffa (1960). The relationship is also known as the 'wage frontier' (Hicks 1965), the 'optimal transformation frontier' (Bruno 1969) and the 'efficiency curve' (Hicks 1973). The duality of the $w - r$ relationship and the $c - g$ relationship, that is, the relationship between the level of consumption output per worker (c) and the rate of growth (g) in steady-state capital theory has been demonstrated by the latter two authors and in more general terms by Burmeister and Kuga (1970); for a detailed account, see Craven (1979).

To begin with, suppose for simplicity that there are only single-product industries with labour as the only primary input and that only one (indecomposable) system of production is known (cf. Sraffa 1960, Part I). Then, with gross outputs of the different products all measured in physical terms and made equal to unity by choice of units and with wages paid at the end of the uniform production period, we have the price system.

$$p = (1 + r)ap + wa_0, \tag{1}$$

where p is the column vector of normal prices, a is the square matrix of material inputs, which is assumed to be productive, and a_0 is the column vector of direct labour inputs. Using the consumption basket d as standard of value or *numéraire*,

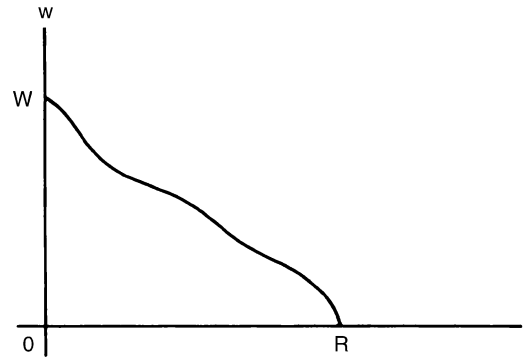
$$dp = 1, \tag{2}$$

we can derive from (1) and (2) the $w - r$ relationship for system (a, a_0)

$$W = \left\{ d[I - (1 + r)a]^{-1} a_0 \right\}^{-1} \tag{3}$$

The relationship is illustrated in Fig. 1. At $r = 0$ the real wage in terms of d is at its maximum value W; it falls monotonically with increases in r, approaching zero as r approaches its maximum value R. (The $w - r$ relationship can be shown to be a straight line if Sraffa's Standard commodity s is used as *numéraire*, where s is a row vector such that $s = (1 + R)\, sa$; cf. Sraffa 1960, chap. IV.)

Let us now assume that several systems are available for the production of the different commodities and that all the production processes exhibit constant returns to scale. We call the set

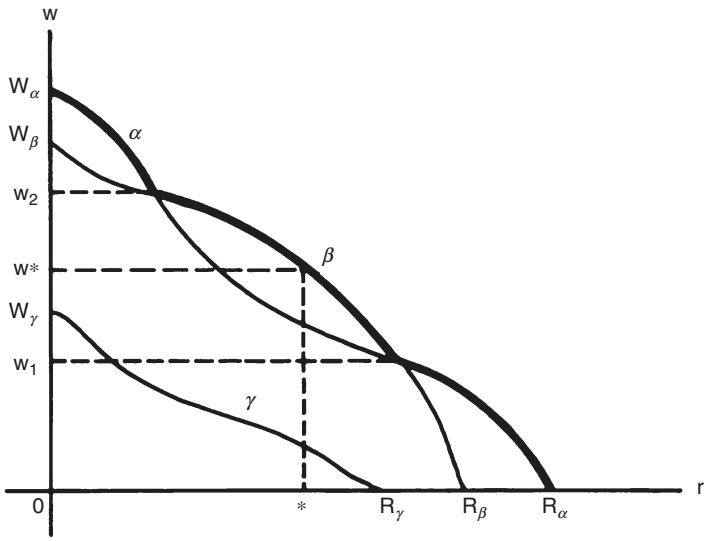


Factor Price Frontier, Fig. 1

of all the alternative methods (or processes) of production known the *technology* of the economic system. From this set a series of alternative *techniques* can be formed by grouping together these methods of production, one for each commodity. Hence there is the question of the *choice of technique*. Under competitive conditions this choice will be exclusively grounded on cheapness, that is, the criterion of choice is that of *cost-minimization*. In the case depicted, it can be shown that the competitive tendency of entrepreneurs to adopt whichever technique is cheapest in the existing price situation, will for a given w (or, alternatively, r) lead to the technique yielding the highest $r(w)$, whereas techniques yielding the same $r(w)$ for the same $w(r)$ are equiprofitable and can co-exist (cf. Garegnani 1970, p. 411).

What has just been said is illustrated in Fig. 2. It is assumed that only three alternative techniques, α β and γ, are available, each of which is represented by the associated $w - r$ relationship; since w is always measured in terms of the consumption basket d, all three relationships can be drawn in the same diagram. Obviously, technique γ is inferior and will not be adopted. Technique α will be chosen for $0 < w < w_1$ and $w_2 < w \leq w_\alpha$, while technique dominates at $w_1 < w < w_2$; there are two *switch points* (at $w = w_1$ and $w = w_2$, respectively) at which both techniques are equiprofitable. The heavy line represents the economy's $w - r$ *frontier* (or 'factor price frontier') and is the outer envelope of the $w - r$ relationships. At a level of the wage rate w^*, for example, technique β will be adopted giving a rate of profit r^*. (For a discussion of more

Factor Price Frontier, Fig. 2



general cases of single production, see Pasinetti 1977, ch. VI; for a reformulation of some results in capital theory in terms of the so-called 'dual' cost and profit functions, see Salvadori and Steedman 1985; on the maximum number of switch points between two production systems, see Bharadwaj 1970.)

Figure 2 shows that the same technique (α) may be costminimizing at more than one level of the wage rate (rate of profit) even though other techniques (here β) dominate at wage rates in between. The implication of this possibility of the *reswitching* of techniques (and of the related possibility of *reverse capital deepening*) is that the direction of change of input proportions cannot be related unambiguously to changes in the distributive variables. This can be demonstrated by making use of the duality between the $w - r$ and the $c - g$ frontier. Denoting the value of net output per labour unit by y and the value of capital per labour unit by k, we have in steady- state equilibrium

$$y = w + rk = c + gk. \tag{4}$$

Solving for k we get

$$k = (c - w)/(r - g) \tag{5}$$

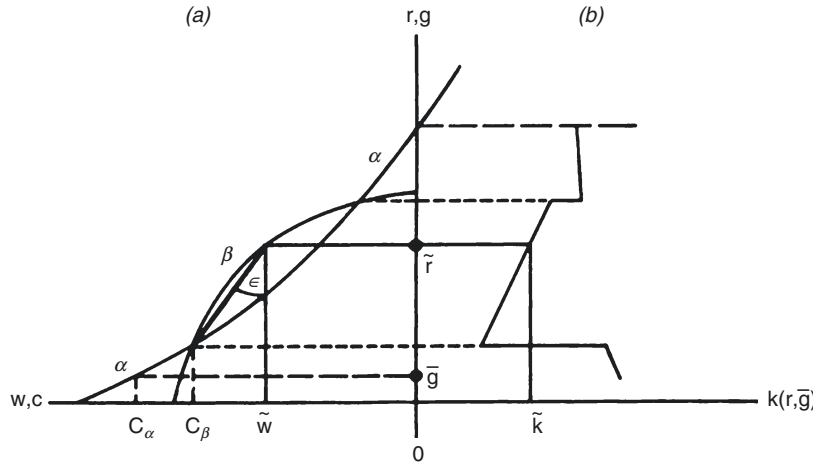
except in golden rule equilibrium ($g = r$), where k can be shown to be (minus) the slope of the golden rule $w - r$ relationship at the going level of r. In Fig. 3(a) the frontier built of two techniques, α and β, is depicted. The rate of growth is fixed at the level $\overline{g}$, to which correspond c_α, and c_β. For values of $r \geq \overline{g}$, that is, on the right side of the golden rule, Fig. 3(b) gives the corresponding value of k $(r,\overline{g})$. For example, at $\tilde{w}$ technique β will be chosen, yielding a rate of profit $\tilde{r}$ the associated capital intensity is given by

$$\tan e = \left(c_\beta - \tilde{w}\right)/(\tilde{r} - \tilde{g}) = \tilde{k}.$$

Figure 3(b) shows that the capital–labour ratio need not be inversely related to the rate of profit as neoclassical long-period theory maintained. In more general terms, it cannot be presumed that input uses, per unit of output, are related to the corresponding 'factor prices' in the conventional way (see Metcalfe and Steedman 1972, and Steedman 1985). This result calls in question the validity of the traditional demand and supply approach to the determination of quantities, prices and income distribution.

The results stated above essentially carry over to the more general case with fixed capital, pure joint production and several primary inputs, such as land and labour of different qualities, provided the formalization of the problem is appropriately adapted to the specific case under consideration. Here it suffices to point out a few additional aspects of the choice of technique problem.

Factor Price Frontier, Fig. 3



With fixed capital there is always such a problem to be solved. This concerns both the choice of the system of operation of plant and equipment, that is, for example, whether a single or a double-shift system is to be adopted; and the choice of the economic lifetimes of fixed capital goods. During the capital theory debates of the 1960s and early 1970s attention focussed on the latter aspect of the use of capital. It was shown that with decreasing or changing efficiency of the durable capital good, cost minimization implies that for a given level of the rate of profit, premature truncation is advantageous as soon as the price (book value) of the partly worn out item becomes negative. While the $w - r$ relationship for a given truncation may slope upwards over some range of r, the $w - r$ frontier consists only of those parts of the $w - r$ relationships that are downward-sloping. Moreover, it was demonstrated that the frontier can display the *return of the same truncation* (cf., for example, Hagemann and Kurz 1976). As to the other aspect of capital utilization, a similar possibility can be shown to exist: the *return of the same system of operation of plant and equipment* (cf. Kurz 1986). Both phenomena are of course variants of the reswitching of techniques.

In systems with pure joint production a choice of technique is inherent, even where the number of processes available does not exceed the number of products. Sraffa's approach to joint production is in terms of 'square' systems of production, that is, systems where the number of processes operated is equal to the number of commodities (i.e. positively-priced products). However, as Salvadori (1982) has shown, in such a framework a cost-minimizing system does not need to exist. A way out of this impasse may be seen in a formalization of joint production that is similar to von Neumann's. In such a formalization the free disposal assumption plays a crucial role. It can be shown that the $w - r$ frontier is downward-sloping, even though individual $w - r$ relationships may have positive ranges.

See Also

- ▶ Reswitching of Technique
- ▶ Sraffian Economics
- ▶ Two-Sector Models

Bibliography

Bharadwaj, K. 1970. On the maximum number of switches between two production systems. *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 106 (December): 409–429.

Bruno, M. 1969. Fundamental duality relations in the pure theory of capital and growth. *Review of Economic Studies* 36 (January): 39–53.

Burmeister, E., and K. Kuga. 1970. The factor price frontier, duality and joint production. *Review of Economic Studies* 37 (January): 11–19.

Craven, J. 1979. Efficiency curves in the theory of capital: A synthesis. In *The measurement of capital, theory and*

practice, ed. K.D. Patterson and K. Schott. London: Macmillan.
Garegnani, P. 1970. Heterogeneous capital, the production function and the theory of distribution. *Review of Economic Studies* 37 (July): 407–436.
Hagemann, H., and H.D. Kurz. 1976. The return of the same truncation period and reswitching of techniques in neo-Austrian and more general models. *Kyklos* 29 (December): 678–708.
Hicks, J.R. 1965. *Capital and growth*. Oxford: Oxford University Press.
Hicks, J.R. 1973. *Capital and time, a neo-Austrian theory.* Oxford: Oxford University Press.
Kurz, H.D. 1986. 'Normal' positions and capital utilisation. *Political Economy* 2 (May): 37–54.
Metcalfe, J.S., and I. Steedman. 1972. Reswitching and primary input use. *Economic Journal* 82 (March): 140–157.
Pasinetti, L.L. 1977. *Lectures on the theory of production*. London: Macmillan.
Ricardo, D. 1971. *The works and correspondence of David Ricardo*, Edited by P. Sraffa in collaboration with M.H. Dobb, vol. VIII. Cambridge: Cambridge University Press.
Salvadori, N. 1982. Existence of cost-minimizing systems within the Sraffa framework. *Zeitschrift für Nationalökonomie* 42 (September): 281–298.
Salvadori, N., and I. Steedman. 1985. Cost functions and produced means of production: Duality and capital theory. *Contributions to Political Economy* 4 (March): 79–90.
Samuelson, P.A. 1957. Wages and interest: A modern dissection of Marxian economic models. *American Economic Review* 47 (December): 884–912.
Samuelson, P.A. 1962. Parable and realism in capital theory: The surrogate production function. *Review of Economic Studies* 29 (June): 193–206.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Steedman, I. 1985. On input 'demand curves'. *Cambridge Journal of Economics* 9 (June): 165–172.

Factor Prices in General Equilibrium

Michael Mandler

Abstract

In general equilibrium models with linear or nonlinear activities, factor prices can be indeterminate and agents will have an incentive to non-competitively manipulate prices even if they are small relative to the market. The indeterminacy cannot occur at generic endowments, but the non-generic endowments where it does occur will arise endogenously as an equilibrium outcome when some factors, such as capital goods, are produced. This endogenous indeterminacy creates a hold-up problem since investors need not earn the rate of return that obtains in an intertemporal competitive equilibrium. Unlike the classical hold-up problem, factor-price indeterminacy is not attributable to there being few agents or bilateral monopoly.

Keywords

Arrow–Debreu model of general-equilibrium; Differentiable production function; Excess demand and supply; Factor prices in general equilibrium; Factor-price indeterminacy; Hold-up; Imperfect competition; Intertemporal efficiency; Intertemporal equilibrium; Leontiev production function; Linear activities model; Nonmarket institutions; Regular economies; Sequential-trading equilibrium; Uniqueness of equilibrium; Walras's Law

JEL Classifications

F11; N70

Introduction

At first glance, the Walrasian general equilibrium model does not offer a theory of factor prices. Factors are goods supplied by agents to firms which then use them to produce outputs. In the general equilibrium model, there is no such class of goods: one and the same good can simultaneously be used as an input by some firms, produced as an output by other firms, sold by some consumers, and purchased and consumed by other consumers. Indeed, the general equilibrium model's abstraction from the minutiae of how particular goods are used is one of the theory's great advantages. For many of the classical concerns of the Walrasian tradition – the existence of equilibrium, optimality – these details are irrelevant.

Even if there is a category of factors that consumers sell and firms buy, it is hard to see any distinctive properties of these goods. While factor supply functions can exhibit perverse responses to price changes, so can output demand functions. The responses of firms to price changes are better behaved, and firm factor demands may seem to be governed by a distinctive principle: a firm's demand for a factor diminishes in its own price while a firm's supply of an output increases in its own price. While correct, these two fundamental rules of producer comparative statics are really reflections of a single law, as Samuelson (1947) showed long ago. Suppose in an ℓ-good economy that a profit-maximizing firm with production set $Y \subset R^{\ell}$ chooses $y = (y_1, \ldots, y_\ell) \in Y$ when facing prices $p = (p_1, \ldots, p_\ell)$ and $\hat{y} \in Y$ when facing $\hat{p}$. Since each decision is profit-maximizing, $p \cdot y \geq p \cdot \hat{y}$ and $\hat{p} \cdot \hat{y} \geq \hat{p} \cdot y$ and hence $(\hat{p} - p) \cdot (\hat{y} - y) \geq 0$. If only one price differs at p compared to $\hat{p}$, say the first, then $(\hat{p}_1 - p_1)(\hat{y}_1 - y_1) \geq 0$.So if $\hat{p}_1 > p_1$ then $\hat{y}_1 \geq y_1$.Both of the comparative statics rules now follow from the appropriate sign restrictions on y_1 and $\hat{y}_1$: when both are positive we conclude that the output of good 1 supplied by the firm must be weakly increasing in its price, while if both are negative we conclude that the factor demand for good 1 must be weakly decreasing in its price (since $\hat{y}_1 \geq y_1$ and $(\hat{y}_1, y_1) \leq 0$ imply $|\hat{y}_1| \leq |y_1|$). It is tempting to conclude that there is no special general equilibrium principle of factor demands, just a specific application that follows when the sign convention for factors is inserted.

Factor-Price Indeterminacy

The demand for and supply of factors can nevertheless exhibit distinctive properties, although they are consistent with the generalities pointed out in the previous section. These properties do not matter for the most of the classical results of general equilibrium theory, but they can undermine one result, the generic determinacy (local uniqueness) of equilibria.

The first distinguishing trait of factors is that sometimes they do not provide any direct utility and are useful only as inputs in production. Consumers will supply to the market their entire endowment of such 'pure' factors and hence supply will be inelastic with respect to price changes. As we will see, what matters is local unresponsiveness to prices. Perhaps when a factor such as iron ore is sufficiently cheap in terms of consumption goods consumers will find some direct use for it and hence have an excess demand that locally varies as a function of prices. But above some minimum price, consumers will not consume any iron ore and in this range consumers' excess demand will be inelastic. Second, technology can restrict the number of ways in which factors can be productively combined. The extreme case occurs with fixed coefficients – the Leontiev production function – where to produce one unit of a good just one combination of factors will do. More flexible is the linear activities model where finitely many constant-return-to-scale techniques are available to produce one or more goods. Factors then may be combined in various configurations but some factor proportions cannot be used productively (that is, without disposing of some of the factors). Nonlinear activities are qualitatively similar but do not require constant returns to scale. In all these cases, production sets have a kinked rather than smooth (differentiable) surface. Consequently factor prices can be adjusted at least slightly from one equilibrium configuration without changing the quantity of factors that profit-maximizing producers will demand when producing a given quantity of output (or vector of outputs). In the Leontiev case, picture the multiple price lines that can support the model's L-shaped isoquants. Of course, production sets do not have to exhibit kinks; for example, they will be smooth when each output is a differentiable function of factor inputs. Any change in relative factor prices will then lead to a change in factor demand.

Factors of production thus are distinctive in that both demand and supply can be unresponsive to certain types of price changes. Factor demand and supply do not have to display this unresponsiveness, but under plausible circumstances permitted by the general equilibrium model they will. Inelastic factor demand and supply in turn can lead to an indeterminacy of factor

prices. For a simple example, suppose an economy has one consumption good, produced by a single linear activity that requires a_1 units of one factor and a_2 units of a second factor to yield one unit of output. Set the price of consumption equal to 1, let w_1 and w_2 be the two factor prices, let the endowments of the two inelastically supplied factors be $e_1 \geq 0$ and $e_2 \geq 0$, and let y be the sole activity usage level. An equilibrium $(w_1, w_2, y) \geq 0$ where the consumption good is produced and has a positive price must satisfy three conditions: (i) $a_1w_1 + a_2w_2 = 1$ (the activity breaks even), (ii) $a_iy \leq e_i$, for $i = 1, 2$ (market-clearing for factors), and (iii) $a_iy < e_i \Rightarrow w_i = 0$ for $i = 1, 2$ (factors in excess supply have a 0 price). On the assumption that the demand for output equals factor income, which is a form of Walras's law, (i)–(iii) imply that the market for output clears. Evidently equilibrium must satisfy $y = \min[e_1/a_1, e_2/a_2]$. By (iii), the two factors will both have a strictly positive price only if

$$\frac{e_1}{a_1} = \frac{e_2}{a_2}, \tag{1}$$

in which case any $w = (w_1, w_2) \geq 0$ that satisfies (i) will be an equilibrium w: indeterminacy therefore obtains when (1) holds. We defer for a little while the question of whether this knife-edge condition is likely to be satisfied.

Fixed coefficients and inelastic factor supply do not always lead to indeterminate factor prices. Prior to the invention of the differentiable production function and for a while thereafter, the standard cure for factor-price indeterminacy was to argue that, even if each industry uses factors in fixed proportions, those proportions will differ across industries; variations in factor prices will then lead to changes in relative output prices, and thus to changes in output demand that feedback to changes in factor demand (Cassel 1924; Wieser 1927). Substitution in consumption can thereby play the same equilibrating role as the technological substitution of inputs in production. For the simplest example, suppose we supplement the above single-sector economy with a new sector that uses b_1 units of the first factor and b_2 units of the second factor to produce one unit of a second consumption good. If we keep the price of the first consumption good equal to 1, and let p_b be the price of the second consumption good, then when both activities break even the equalities

$$a_1w_1 + a_2w_2 = 1, \quad b_1w_1 + b_2w_2 = p_b$$

must be satisfied. As long as $a_1/a_2 \neq b_1/b_2$, it will not be possible to adjust w without also changing the relative price of the consumption goods p_b. When w_1/w_2 increases, the consumption good that uses factor 1 more intensively will rise in price, presumably diminishing demand for that good and thus diminishing the demand for factor 1. Even if demand for consumption is a perverse function of prices, this two output–two factor model will still typically have determinate prices as long as both activities break even.

A general linear activity analysis model will clarify when the determinate and indeterminate cases arise. The linearity of the activities serves only to simplify the model's equilibrium conditions. There will be two types of goods: factors, which give no utility and are inelastically supplied, and consumption goods, which do give utility. Despite their name, consumption goods can be used as inputs and nonproducible but they must provide utility to some agents. We now adopt the standard sign convention and define an activity to be a vector, with as many coordinates as there are commodities, whose positive coordinates give the quantities of goods produced and negative coordinates give the quantities of goods used when the activity is operated at the unit level. In equilibrium the excess demand for each good must be non-positive, each good in excess supply must have a 0 price, each activity must earn non-positive profits, and each activity in use must earn 0 profits. Since determinacy and indeterminacy are purely local events, a search for equilibrium prices and activity near a reference equilibrium can ignore the 'slack' equilibrium conditions, the market-clearing condition for any good in excess supply and the no-positive-profits condition for any activity that either makes strictly negative profits or utilizes and produces only goods in excess supply: for small adjustments of

prices and activity levels, the excluded goods will remain in excess supply and the excluded activities will continue to make negative profits or continue to use and produce only goods in excess supply (and hence continue to break even). Call any good not in excess supply and any activity that breaks even and that uses or produces at least one good not in excess supply 'operative'. Given some reference equilibrium with ℓ operative consumption goods, m operative factors, and n operative activities, let A be the $(\ell + m) \times n$ activity analysis matrix whose rows and columns correspond to the operative goods and activities, let y be the n-vector of operative activity levels, let p be the ℓ-vector of prices for the operative consumption goods, let w be the m-vector of prices for the operative factors, let $z(p, w)$ be the excess demand function for the operative consumption goods, which we assume is homogeneous of degree 0 in (p, w), and finally let e be the m-vector of inelastic supplies of the operative factors. Walras's law then states that $p \cdot z(p, w) = w \cdot e$. Equilibria $(p, w, y) \geq 0$ are locally characterized by the equalities

$$(z(p,w), -e) = Ay, \quad (p,w)'A = 0. \tag{2}$$

(All vectors are column vectors and $'$ denotes transposition.) Bear in mind that the market-clearing and no-positive-profit inequalities excluded from (2) vary by equilibrium; the activities and goods operative in one equilibrium need not be operative in another. We assume henceforth that, at any equilibirum, each of the operative activities is used at a strictly positive level and that each operative good has a strictly positive price, $(p, w, \underline{y}) \gg 0$. As usual, the homogeneity of demand allows us to set one of the positively priced goods to be the numéraire and Walras's law implies that one of market-clearing conditions is redundant. So we set the price of the first consumption good not in excess supply to equal 1 and put aside the market-clearing condition for this good. Letting $\overline{z}(p,w)$ denote $z(p, w)$ without the first coordinate, $\overline{A}$ denote A without the first row, and $\overline{p}$ denote p with the first coordinate set equal to 1, (2) can be written

$$(\overline{z}(\overline{p},w), -e) = \overline{A}y \tag{3}$$

$$(\overline{p},w)'A = 0. \tag{4}$$

Any small change in $(\overline{p}, w, y)$ that satisfies (3) and (4) will then continue be an equilibrium: the variables $(\overline{p}, w, y)$ will remain positive, all excluded goods will remain in excess supply, and all excluded activities will continue to make negative profits or continue to use and produce only goods in excess supply.

The most conspicuous case of factor-price indeterminacy occurs when $m > n$, that is, when there are more operative factors than operative activities. If, beginning at some reference equilibrium, we fix y at its equilibrium value, then as $(\overline{p}, w)$ varies the market-clearing conditions for factors in (3) will continue to be satisfied. But the remaining equilibrium conditions – (4) and the market-clearing conditions for consumption goods in (3) – comprise $n + \ell - 1$ equations in the $\ell - 1 + m$ variables $(\overline{p}, w)$. Hence, if $m > n$ and as long these remaining equilibrium conditions satisfy a rank condition, which allows the implicit function theorem to be applied, indeterminacy will occur. The economy considered earlier where two factors are used by one activity qualifies as an example of the $m > n$ type of indeterminacy, while the economy where two factors are used to produce two goods does not.

A slight variation of this argument applies to a subset of factors. Suppose that $\hat{m}$ of the m operative factors are used by only $\hat{n}$ of the n operative activities, and that $\hat{m} > \hat{n}$. Thus the remaining $n - \hat{n}$ operative activities have 0 entries in the rows of A that correspond to these $\hat{m}$ factors. If we fix the $\hat{n}$ coordinates of y for the activities that do use these $\hat{m}$ factors, then, as the remaining endogenous variables (the other $n - \hat{n}$ activity levels, $\overline{p}$, and w) change, the market-clearing conditions for the $\hat{m}$ factors will continue to be satisfied. Moreover, the number of remaining endogenous variables is $n - \hat{n} + \ell - 1 + m$ while the number of remaining equilibrium conditions is $\ell - 1 + m - \hat{m} + n$. The difference between the number of remaining variables and remaining equilibrium conditions is therefore $\hat{m} - \hat{n}$ and so there are more variables

than equilibrium conditions. Indeterminacy therefore obtains (again, given a rank condition).

Factor-price indeterminacy, whether for an economy as a whole or for a subset of an economy's factors, depends critically on production sets that exhibit kinks. By fixing a set of activity levels, the above indeterminacy argument fixes a vector of factor demands and finds a multiplicity of prices at which firms will demand exactly those quantities. If the aggregate production set were smooth, a fixed vector of firm factor demands would be supported by only one vector of relative factor prices.

Factor-price indeterminacy brings dramatic behavioural consequences: agents have a strong incentive to manipulate factor prices and hence markets cannot function competitively. In the two factor–one activity example, where the endowments satisfy (1), the tiniest withdrawal of either factor $i = 1$ or $i = 2$ from the market will lead the other factor to be in excess supply and have price 0 and hence cause factor i's price to jump to $1/a_i$. No matter how small an owner of factor i is as a proportion of the market, it will be in his or her interest to remove a small amount of i from the market. Agents therefore will not behave like price-takers. When more activities are present, the jump in factor prices need not be as large, but a jump will still occur for an arbitrarily small withdrawal of a factor, and hence the incentive to manipulate will remain. The distinctive mathematical feature of factor-price indeterminacy that drives this conclusion is that the equilibrium correspondence fails to be lower hemicontinuous. (The equilibrium correspondence is the correspondence from the parameters of the model, such as the endowments e, to the endogenous variables $(\overline{p}, w, y)$.) When the endowments of factors lead to an indeterminate equilibrium, it will usually be impossible at nearby endowment levels to find equilibrium prices near to the prices of the indeterminate equilibrium. Other varieties of indeterminacy in the general equilibrium model, such as the indeterminacy of the overlapping generations model, do not suffer from such a failure of lower hemicontinuity and therefore do not invite market manipulation (see Mandler 2002).

The Emergence of Factor-Price Indeterminacy Through Time

We saw in the two factor–one activity example that indeterminacy occurs only if a knife-edge condition on endowments is satisfied. This observation applies to the broader species of factor-price indeterminacy as well. Suppose again that at some reference equilibrium $\hat{m}$ operative factors are used by $\hat{n} < \hat{m}$ operative activities, let $\hat{e}$ be the endowments of these $\hat{m}$ factors, let $\hat{y}$ be the activity levels for the $\hat{n}$ activities, and let $\hat{A}$ be the $\hat{m} \times \hat{n}$ submatrix of A formed by the rows for the $\hat{m}$ factors and the columns for the $\hat{n}$ activities. Then $\hat{A}\hat{y} = \hat{e}$. But since $\hat{A}$ has more rows than columns, for almost every value of $\hat{e}$, $\hat{A}\hat{y} = \hat{e}$ will have no solution. Hence, for most levels of an economy's endowments, there will be no equilibrium at which $\hat{m}$ operative factors are used by fewer than $\hat{m}$ operative activities. While the failure in these so-called generic cases of the indeterminacy arguments we have given does not show that equilibria are generically locally unique, the literature on *regular economies* (see in particular Mas-Colell 1975, 1985; Kehoe 1980, 1982) has shown that, for generic endowments and preferences, general equilibrium models with linear or nonlinear production activities do have locally unique equilibria.

The determinacy qst, however, does not end here. An economy's endowments of produced inputs – capital goods – are in any long-term view endogenous variables not parameters. Consequently, even though factor-price indeterminacy does not arise for generic endowments, it is conceivable that those special endowments that lead to indeterminacy will systematically arise as the equilibrium activity of an economy unfolds through time. To see that this can indeed happen, we partition an intertemporal economy's dates into two periods, a first period where goods are either consumed or invested in the production of factors, and a second period where the factors produced by first-period activities and natural endowments are used to create consumption goods (possibly also with the aid of intermediate inputs produced within the second period). To test whether the nongeneric factor endowments that

lead to indeterminacy are likely to appear, we consider intertemporal economies where the endogenous equilibrium production of second-period factors leads the total stock of these factors to assume the nongeneric values where indeterminacy arises. If this endogenous second-period indeterminacy obtains for a robust family of equilibria (the equilibria of a nonempty open set of economies), then *sequential indeterminacy* occurs (Mandler 1995).

In the Arrow–Debreu view of an intertemporal economy, agents trade just once at the beginning of economic time; after these initial contracts are signed, no further trade occurs, goods are just delivered. To allow for trade at multiple dates, and thus give indeterminacy in later time periods a chance to appear, we assume instead that agents transfer wealth between periods by borrowing or lending assets. Agents then will typically trade every period, and the economies that appear in later periods will have endowments that are endogenously determined by trade in the initial periods. Moreover any indeterminacy of prices in later periods will change the quantities of goods exchanged and hence change agents' utilities. In our setting, with just two periods, we can let the activities that produce second-period factors serve as assets: agents in the first period will buy or sell rights to the outputs of the activities that produce the second-period factors and then in the second period receive or deliver the second-period factors they contracted for in the first period and use their income to trade for consumption. The allocation achieved by a two-period Arrow–Debreu intertemporal equilibrium will occur in an equilibrium with two sequential periods of trade if (*a*) agents in the first period unanimously anticipate a second-period price vector, (*b*) given those expectations, goods and asset markets in the first period clear, and (*c*) given asset deliveries, second-period markets clear at the anticipated prices. We omit the routine details of how to decompose an intertemporal equilibrium into a sequential-trading equilibrium (see Radner 1972) and will just write one equilibrium condition explicitly, the market-clearing equality for second-period factors.

As usual, we consider some reference equilibrium and ignore those goods in excess supply and those activities that make strictly negative profits or that use and produce only goods in excess supply. If there are k operative goods in period 1, and ℓ operative consumption goods and m operative factors in period 2, the activity analysis matrix for the operative goods and activities takes the form

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_{c2} \\ A_{f1} & A_{f2} \end{pmatrix} \begin{matrix} k \\ \ell \\ m \end{matrix}$$

where the subscript c or f indicates whether the rows are for consumption goods or factors and the subscript 1 or 2 indicates the time activities begin operation. Since presumably the second-period factors are the outputs of time 1 activities and the inputs of time 2 activities, it makes sense to suppose $A_{f1} \geq 0$ and $A_{f2} \leq 0$. If we let y_i denote the activity levels for operative activities that begin in period i and e the endowment of operative second-period factors, the market-clearing equality for operative second-period factors is

$$A_{f1}y_1 + A_{f2}y_2 + e = 0. \tag{5}$$

In the background lie the remaining equilibrium conditions: market-clearing conditions for excess-supply factors and for all consumption goods, and nonpositive profit conditions for activities.

Consider the restrictions that (5) places on the number of operative factors. If the number of operative activities in the two periods that produce or use the m operative second-period factors is less than m, then, for almost every e, (5) will have no solution $y = (y_1, y_2) \geq 0$. Similarly if there is a subset of $\hat{m}$ operative second-period factors where the number of operative activities in the two periods that produce or use these factors is less than $\hat{m}$, then again (5) will usually have no solution. We may therefore dismiss these cases as unlikely, in line with the literature on regular economies. In the remaining cases, where for each subset of $\hat{m}$ operative second-period factors the number of operative activities in the two periods that produce or use these factors is greater than or equal to $\hat{m}$, then (5) can have a solution $y \geq 0$ for a

robust (open) choice of endowment levels e. But in these latter cases it could well be that some subset of operative second-period factors – say the entire set of all m of these factors – is used by fewer than m operative *second-period* activities. For an example, let $m = 2$, suppose that the first factor has no endowment but is produced by an activity with factor output coefficient c_1 while the second factor has a positive endowment in the second period and is not produced. In the second period, both factors are used by one activity with factor usage coefficients a_1 and a_2. Then (5) consists of the two equalities

$$\begin{aligned} c_1 y_1 + a_1 y_2 &= 0, \\ a_2 y_2 + e_2 &= 0. \end{aligned} \tag{6}$$

Evidently if $a_1 < 0$, $a_2 < 0$, $c_1 > 0$, and $e_1 > 0$, then a solution $y \gg 0$ to (6) exists and is robust: for a small variation in the production coefficients or the endowment, a solution $y \gg 0$ will continue to exist. In this equilibrium, factor 2 is produced in just the quantity necessary to ensure that neither factor 1 nor factor 2 is in excess supply. For a second example, suppose that factor 2 is produced as well and also has no endowment, and let y_{1i} denote the usage level of the activity that produces factor i. Then (6) is replaced by $c_1 y_{11} + a_1 y_2 = 0$ and $c_2 y_{12} + a_2 y_2 = 0$. Now efficiency and hence equilibrium will usually *require* that the two factors are produced in quantities that leave neither in excess supply in period 2; if, say, factor 1 were in excess supply and if y_{11} could be lowered, thereby increasing the output of some first-period consumption good, an inefficiency would exist, which is impossible in equilibrium.

Once agents arrive at period 2, they trade again but now the factor outputs produced by the activities that began in period 1 are exogenously given. So in the example given by (6) the endowment of factor 1 in period 2 equals $c_1 y_1$ and one may readily check that this quantity along with e_2 of factor 2 satisfies the knife-edge condition (1). Thus, despite seeming to be unlikely at a given point in time, the endowments that lead to indeterminacy can endogenously arise.

Intertemporal general equilibrium economies therefore can be sequentially indeterminate. Moreover, factor-price indeterminacy is typically the only source of endogenous indeterminacy. Let us call the equilibria that occur in the later periods of operation of a sequential-trading equilibrium and that confirm the expectations formed in the initial period 'continuation equilibria'. A continuation equilibrium is *indeterminate* if it sits amid a continuum of other (usually non-continuation) equilibria.

Sequential indeterminacy th (Mandler 1995). For a generic set of intertemporal economies with linear activities, a continuation equilibrium is indeterminate at some date t if and only if there is a set of $\hat{m}$ operative factors appearing at t or later that are used or produced by fewer than $\hat{m}$ operative activities that begin at t or later.

In contrast, when production sets are smooth, endogenous endowments do not lead to indeterminacy; typically continuation equilibria are locally unique (Mandler 1997).

Factor Price Indeterminacy and the Hold-Up Problem

The endogenous factor-price indeterminacy of the previous section is *not* an indeterminacy of the equilibria of the entire intertemporal economy or of the corresponding sequential-trading equilibria. As long as the non-produced endowments of every period of an intertemporal economy avoid certain nongeneric values, and barring flukes in preference or technology coefficients, only a finite number of intertemporal equilibria will exist. It follows that in a two-period model that displays sequential indeterminacy, almost all of the infinite multiplicity of equilibria of the second-period economy could not form part of a two-period sequential-trading equilibrium: if the prices of almost any of the second-period equilibria were anticipated in period 1, they would be inconsistent with market clearing. Specifically, if anticipated second-period prices were to vary slightly from the values that hold in a sequential-trading equilibrium, then either assets would no longer share the same rate of return or the common rate of return on assets would change, and hence typically markets would not clear. But bygones are bygones: once period

1 is past, even the second-period equilibria that violate the requirements of an intertemporal equilibrium are equilibria nonetheless when the economy arrives at its second period.

Moreover second-period indeterminacy will prevent sequential-trading equilibria from proceeding smoothly through time: they will be virtually certain to unravel. Since factor prices are indeterminate in the second period, rational agents will predict that an investment in an activity producing a second-period factor will not except by chance earn the rate of return anticipated in the first period of a sequential-trading equilibrium. Investments will therefore differ from their Walrasian levels. The predictions of the general equilibrium model thus become untenable when agents trade repeatedly through time and factor-price indeterminacy is present, even though all the classical presuppositions of the model – price-taking agents, no distortions and so on – obtain.

The inability of second-period markets to ensure that assets earn the rate of return necessary for efficiency amounts to a hold-up problem, but the cause of the problem differs from the conventional diagnosis. In the classical hold-up problem, the owners of two complementary factors Nash bargain over the revenue they jointly earn; hence, if the owner of one of the factors invests to improve the quality of his factor, the owner recoups only a fraction of the increment to revenue, and consequently investment is inefficiently low (Hart 1995). The problem, it would seem, is that the factor owners form a bilateral monopoly and cannot purchase each other's services on a competitive market. What we have seen, however, is that a hold-up problem can arise with perfectly competitive markets. Even if factor owners can purchase all complementary factors on competitive markets, factor-price indeterminacy can prevent investments in factors from earning the rate of return required in intertemporal equilibrium (and hence the rate necessary for efficiency): an unguided market has no means to select from the continuum of equilibrium factor prices the specific prices that deliver intertemporal efficiency. Factor markets moreover will not operate competitively in the presence of factor-price indeterminacy, which is another cause for the rate of return to deviate from its competitive equilibrium value. For both reasons, the efficient Walrasian levels of investment need not occur.

Just as in the classical hold-up problem, long-term contracts can mitigate the troubles that factor-price indeterminacy brings. If labour is among the factors in an economy displaying factor-price indeterminacy, then a labour contract may be able to force trading at prices that allow intertemporal efficiency and prevent labourers or capital goods owners from manipulating factor prices by withdrawing their services from the market. Of course, as in the classical hold-up problem, the incompleteness of contracts may hamper the ability of this solution to deliver first-best efficiency. Alternatively, when a set of complementary factors displays factor-price indeterminacy and consists solely of produced goods, then a bundling of the complementary factors in an asset portfolio – that is, in a 'firm' – can eliminate the incentive to manipulate prices. From the vantage point of factor-price indeterminacy, unions and labour contracts and the firm as an institution emerge as devices to enforce competitive equilibria, not as consequences of imperfect competition in factor markets.

Conclusion: Factor-Price Indeterminacy Past and Present

Prior to the Arrow–Debreu transformation of general-equilibrium theory, economists were well aware that linear activities could lead to an indeterminacy of factor prices. The problem was considered from a long-run perspective: a change in a factor price was presumed to persist for many periods, and, although such a change might not lead to an instantaneous change in either the supply or demand for the factor, arguments were deployed for why demand and supply responses would eventually kick in. For example, in response to a wage increase, although existing capital equipment might have fixed labour requirements, newly constructed capital equipment could be built to use labour less intensively. In addition, a wage increase would eventually lead the price of labour-intensive consumption goods to rise, diminishing the demand for these

goods and therefore ultimately for labour as well. This effect does not operate immediately since a wage increase will lead to an offsetting fall in the prices of existing stocks of complementary capital inputs. But the prices of newly produced capital inputs are constrained by break-even requirements; hence, given enough time, the prices of labour-intensive consumption goods will increase. (Robertson 1931, and Hicks 1932, offered the most detailed long-run theories. See Mandler 1999, ch. 2.) Although pre-modern explanations of factor prices faced the indeterminacy problem explicitly, and marshalled a rich array of counter-arguments for why the problem normally will not be severe, the long-run perspective had its drawbacks: the attention to persistent changes in factor prices masked an inability to explain why factor prices cannot temporarily change. The older long-run theories simply assumed that, in the absence of demand or supply shocks, factor prices will be maintained at their long-run equilibrium values. This presumption amounts to a rudimentary version of the rule that in an intertemporal equilibrium prices should fulfill the expectations that agents formed in earlier periods. As we have seen, the market mechanism will not enforce this rule; a supplementary theory of contracts and institutions is necessary. The Arrow–Debreu treatment of factors (and other goods) at different dates as fully distinct goods naturally raises the question of whether prices can deviate from previously anticipated values even in the absence of shocks, and curiously, therefore, the Arrow–Debreu account of markets points to the need for a theory of non-market institutions. Unfortunately, the Arrow–Debreu tradition also took the model of trading at a single point in time as its benchmark. It is only with the combination of goods rigorously distinguished by date, sequential trading, and production sets with kinks that factor-price indeterminacy will appear.

See Also

- ▶ Determinacy and Indeterminacy of Equilibria
- ▶ General Equilibrium
- ▶ Hold-Up Problem

Bibliography

Cassel, G. 1924. *The theory of social economy*. New York: Harcourt.

Hart, O. 1995. *Firms, contracts, and financial structure*. New York: Oxford University Press.

Hicks, J. 1932. Marginal productivity and the principle of variation. *Economica* 25: 79–88.

Kehoe, T. 1980. An index theorem for general equilibrium models with production. *Econometrica* 48: 1211–1232.

Kehoe, T. 1982. Regular production economies. *Journal of Mathematical Economics* 10: 147–147.

Mandler, M. 1995. Sequential indeterminacy in production economies. *Journal of Economic Theory* 66: 406–436.

Mandler, M. 1997. Sequential regularity in smooth production economies. *Journal of Mathematical Economics* 27: 487–504.

Mandler, M. 1999. *Dilemmas in economic theory*. New York: Oxford University Press.

Mandler, M. 2002. Classical and neoclassical indeterminacy in one-shot vs. ongoing equilibria. *Metroeconomica* 53: 203–222.

Mas-Colell, A. 1975. On the continuity of equilibrium prices in constant returns production economies. *Journal of Mathematical Economics* 2: 21–33.

Mas-Colell, A. 1985. *The theory of general economic equilibrium: A differentiable approach*. Cambridge: Cambridge University Press.

Radner, R. 1972. Existence of equilibrium of plans, prices, and price expectations in a sequence of markets. *Econometrica* 40: 289–303.

Robertson, D. 1931. Wage grumbles. In *Economic fragments*. London: P.H. King.

Samuelson, P. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Wieser, F. 1927. *Social economics*. New York: Greenberg.

Fair Allocation

William Thomson

Abstract

We survey the theory of equity in a variety of concretely specified resource allocation models: classical economies with private goods, economies with production, economies with indivisible goods, when monetary compensations are feasible and when they are not, economies with single-peaked preferences, and economies in which the dividend is a

non-homogeneous continuum. We present the central fairness punctual notions, no-envy, egalitarian-equivalence, concepts of equal or equivalent opportunities and the relational principles of monotonicity and consistency.

Keywords

Consistency; Convexity; Efficient allocation; Egalitarian-equivalent allocation; Envy-free allocation; Equality of opportunity; Fair allocation; Indivisible goods; Monotonicity; Shapley value; Single-peaked preferences; Solidarity

JEL Classifications

D2

We survey the theory of equity in concretely specified economic environments. The literature concerns the existence of allocation rules satisfying various requirements of fairness expressed in terms of resources and opportunities understood in their physical sense (and not in terms of abstract entities such as utilities or functionings). For lack of space, we often give only representative references. Detailed treatments of the subject are Young (1994), Brams and Taylor (1996), Moulin (1995, 2003), and Thomson (1995b, 2006c).

Concepts

We introduce concepts central to the classical problem of fair division. These have much broader applicability, but for other models they sometimes have to be reformulated. Also, as models vary in their mathematical structures, the implications of a given concept may differ significantly from one to the other.

In an *economy*, there is a social endowment of resources to be distributed among a group of agents who are collectively entitled to them. For what we call a classical problem of fair division, the resources are infinitely divisible private goods, and preferences are continuous, usually monotonic (sometimes strictly so), and convex. In an *economy with individual endowments*, each agent starts out with a share of society's resources; the issue in this case is to redistribute endowments. In a *generalized economy*, some resources are initially owned collectively and others are individual endowments (Thomson 1992; Dagan 1995). A *solution* associates with each economy a non-empty subset of its set of feasible allocations. A *rule* is a single-valued solution.

An axiomatic study begins with the formulation of requirements on solutions (or rules). Their logical relations are clarified and their implications, when imposed in various combinations, are explored. For each combination of the requirements, do solutions exist that satisfy all of them? If the answer is 'yes', can one characterize the class of admissible solutions?

A *punctual requirement* applies to each economy separately. The main question then is the existence, for each economy in the domain under consideration, of allocations satisfying the requirement. First are bounds on welfares defined agent-by-agent, in an intra-personal way. Some are lower bounds, offering agents welfare guarantees. Others are upper bounds, specifying ceilings on their welfares. An allocation satisfies *no-domination of, or by, equal division*, if no agent receives a bundle that contains at least as much as an equal share of the social endowment of each good, and more than an equal share of the social endowment of at least one good, or a bundle that contains at most as much as an equal share of the social endowment of each good, and less than an equal share of the social endowment of at least one good (Thomson 1995b). It satisfies the *equal-division lower bound* if each agent finds his bundle at least as desirable as equal division (Kolm 1972; Pazner 1977; and many others).

Second are requirements based on interpersonal comparisons of bundles, or more generally, 'opportunities', involving exchanges of, or other operations performed on, these objects. An allocation satisfies *no domination across agents* if no agent receives at least as much of all goods as, and more of at least one good than, some other agent (Thomson 1983a). It satisfies *no-envy* if each agent finds his bundle at least as desirable as that of each other agent (Foley 1967; Kolm 1973, proposes a definition that encompasses many

variants of the concept). The final definition is quite different in spirit: an allocation is *egalitarian-equivalent* if there is a reference bundle that each agent finds indifferent to his own bundle (Pazner and Schmeidler 1978). Given a direction r in commodity space, it is *r-egalitarian-equivalent* if it is egalitarian-equivalent with a reference bundle proportional to r. Of particular interest is when r is the social endowment.

A *relational requirement* prescribes how a rule should respond to changes in some parameter(s) of the economy. The idea of solidarity is central: if the environment changes, and whether or not the change is desirable, but no one in particular is responsible for the change, that is, no one deserves any credit or blame for it (or no one in a particular group of agents is responsible for the change), the welfares of all agents (or all agents in this particular group), should be affected in the same direction: all 'relevant' agents should end up at least as well off as they were initially, or they should all end up at most as well off. In implementing this idea, the focus is usually on a particular parameter. When the parameter belongs to a space that has an order structure, as is frequent, one can speak of the parameter being given a 'greater' or 'smaller' value in that order. Then, together with efficiency, the solidarity idea often implies a specific direction in which welfares should be affected: when a Pareto improvement is possible, all relevant agents should end up at least as well off as they were initially; otherwise, all should end up at most as well off. Thus, solidarity takes the form of a 'monotonicity' requirement. Examples are *resource monotonicity*: if the social endowment increases, all agents should end up at least as well off as they were initially (Thomson 1978; Roemer 1986a, b; Chun and Thomson 1988); *technology monotonicity*, a similar requirement when technology expands (Roemer 1986a; Moulin and Roemer 1989); *population monotonicity*: if population expands, all agents initially present should end up at most as well off as they were initially (Thomson 1983b; Chichilnisky and Thomson 1987; for a survey, see Thomson 2006a).

When the parameter that varies does not belong to a space equipped with an order structure, solidarity retains its general form. For example, *welfare domination under preference replacement* says that if the preferences of some agents change, all agents whose preferences have not changed should end up at least as well off as they were initially, or that all should all end up at most as well off (Moulin 1987b; for a survey, see Thomson 1999). Whether or not the parameter belongs to a space with an order structure, one can imagine simply replacing the initial value taken by the parameter with another value (to which, if there is an order, it may or may not be comparable in the order), and still require that the welfares of all relevant agents should be affected in the same direction.

Another application of the idea is to situations where some agents leave with the resources assigned to them. The requirement that the welfares of all remaining agents should be affected in the same direction, when imposed on efficient rules, often means that these agents should be assigned the same bundles as initially. It can be expressed more simply as *consistency*: given an allocation chosen by a solution for some economy, let us imagine the departure of some agents with their components of it. In the resulting 'reduced economy', the remaining agents should receive the same bundles as initially (for a survey, see Thomson 2006b).

Requirements relative to private endowments, when such exist, may be imposed on rules. For instance, the *individual-endowments lower bound* is the punctual requirement that each agent should end up with a bundle that he finds at least as desirable as his endowment; *individual-endowment monotonicity* is the relational requirement that if an agent's endowment increases, he should end up with a bundle that he finds at least as desirable as the one he got initially (Aumann and Peleg 1974).

Logical relations*; existence*. Under standard assumptions, efficient allocations meeting the equal-division lower bound exist, and so do envy-free and efficient allocations. If preferences are strictly monotonic, no envy implies no-domination. An allocation meeting the equal-division lower bound is not necessarily envy-free. *Equal-division Walrasian allocations* are both envy-free and efficient, and under standard assumptions, they exist. In an economy with an infinite population of agents modelled as a

continuum, and if preferences are sufficiently diverse, a partial converse holds: if an allocation is envy-free and efficient, it is an equal-income Walrasian allocation (Varian 1974; Kleinberg 1980; Champsaur and Laroque 1981; Mas-Colell 1987; Zhou 1992). If preferences are not convex, the existence of envy-free and efficient allocations can be derived from certain assumptions about the structure of the efficient set itself (Varian 1974; Svensson 1983a, 1994b; Diamantaras 1992). In the absence of such assumptions, efficient allocations satisfying no envy, even no domination, may not exist (Maniquet 1999).

An egalitarian-equivalent and efficient allocation may violate no domination. When $|N| > 2$, and if the reference bundle is proportional to the social endowment, then obviously the equal-division lower bound is met, although not no domination. In fact, for some economies, all egalitarian-equivalent and efficient allocations violate no domination (Daniel 1978). An equal-division Walrasian allocation may not be egalitarian-equivalent.

The existence of *r*-egalitarian-equivalent and efficient allocations holds under weak assumptions (Pazner and Schmeidler 1978; Sprumont and Zhou 1999, offer a proof for economies with a continuum of agents).

Variants: extensions. Some solutions are based on comparing across agents the number of agents whom each agent envies and the number of agents who envy him. Envy is *balanced* if, for each agent, these two numbers are equal (Daniel 1975). The existence of allocations with balanced envy holds more generally than is common for other concepts. Other natural ideas are to require of an allocation that all agents should envy the same number of agents, or that all agents should be envied by the same number of agents. But neither definition will do, as soon as efficiency is imposed, because in any economy whose set of feasible allocations is closed under permutations, at an efficient allocation, at least one agent envies no one, and at least one agent is envied by no one (Varian 1974; Feldman and Kirman 1974).

Selections. When envy-free allocations exist, there may be a large number of them and the question of selection arises. A variety of proposals have been made. Some are based on quantifying the extent to which the no envy constraints are exceeded. Conversely, when envy-free allocations do not exist, the extent to which they are violated can also be measured. Measures based on counts of envy relations, or on the adjustments in commodity bundles required to eliminate envy have been proposed (Feldman and Kirman 1974; Varian 1976; Chaudhuri 1985, 1986; Diamantaras and Thomson 1990; Kolpin 1991a). These operations can be adapted so as to extend, or select from, other equity notions, and in a second step, rankings of allocations can be derived (Chaudhuri 1986; Thomson 1995c).

Group fairness. Most of the concepts of the previous pages can be applied to compare the welfares of groups of agents. Central among them are the *equal-division core*, whose definition is straightforward, and *group no envy*: no group should be able to improve the welfares of all of its members if given access to the resources assigned to some other group of the same size. The definition can be adapted to handle groups of different sizes (Kolm 1972; Feldman and Kirman 1974; Green 1972; Khan and Polemarchakis 1978). Under replication, there is a sense in which the set of efficient allocations that are group envy-free converges to the set of equal-division Walrasian allocations (Varian 1974; Kolpin 1991b).

Fairness of trades. The concepts formulated above for allocations can be adapted in various ways to assess the fairness of individual trades when agents are individually endowed (Kolm 1972; Schmeidler and Vind 1972), and to assess the fairness of the trades of groups (Jaskold-Gabszewicz 1975; Yannelis 1983).

Walrasian trades satisfy most of the definitions that have been proposed and under weak assumptions on preferences, for several of the definitions, a converse inclusion holds (Schmeidler and Vind 1972; Shitovitz 1992).

Interesting conceptual issues arise in relating the fairness of allocations and the fairness of trades (Goldman and Sussangkarn 1980; Thomson 1983a).

Economies with Production

A fundamental issue is fair allocation when agents have contributed differently to production, because they have supplied unequal amounts of their time or because they are unequally productive.

A first way to extend the notion of an envy-free allocation to such situations is by having each agent $i \in N$ compare his *complete* consumption bundle (including his consumption of leisure) to those of the other agents. Unfortunately, and even if preferences are quite well-behaved, envy-free and efficient allocations may not exist (Pazner and Schmeidler 1974). Limited exceptions are when all abilities or all preferences are the same (Varian 1974). Another exception is in the two-good case, under a 'single-crossing' assumption on preferences and when the technology is linear (Piketty 1994).

Egalitarian-equivalent and efficient allocations exist quite generally (Pazner and Schmeidler 1978). Also, under appropriate convexity assumptions, existence still holds if the reference bundle in the definition of egalitarian-equivalence is required to be proportional to the average consumption bundle.

An alternative proposal is to recognize the envy of agent $j \in N$ by agent $i \in N$ only after agent i's consumption of leisure is adjusted for him to produce what agent j produces (Varian 1976; Otsuki 1980). The concept is well defined only if the production set is additive. A proof of the existence of such *productivity-adjusted envy-free* and efficient allocations can be given along the lines of the 'Walrasian' proof of existence of envy-free and efficient allocations in exchange economies and under similar assumptions (Varian 1974). Some have objected to the definition because it lets agents with high productivity appropriate the benefits of their greater skills. Alternative concepts have been defined that attempt to distribute across agents these benefits (Pazner and Schmeidler 1978; Varian 1974; Pazner 1977). The main proposal here has been to take advantage of the instrumental value of the Walrasian solution in delivering envy-free allocations when there is no production and in providing equal opportunities: here, one operates the Walrasian solution from equal division of all goods, including time endowments. Svensson (1994b) states an existence result for allocations at which implicit incomes are equal.

Non-convexities in technologies present another difficulty for the existence of envy-free and efficient allocations (Vohra 1992). Vohra proposes to weaken no envy by imposing a certain symmetry among all agents with respect to possible occurrences of envy (see also Varian 1974). Existence holds without any convexity assumption on either preferences or technologies. A critical one, however, is that there be no agent-specific input (Vohra 1992).

Next, we turn to criteria that, by contrast with the previous ones, can be evaluated agent-by-agent, just like the equal-division lower bound. First, for each agent, we imagine an economy composed of agents having the same preferences as his, and we identify their common welfare under efficiency and *equal treatment of equals*. We take this welfare as a bound, thereby defining the *identical-preferences lower bound*. For nowhere-increasing returns-to-scale, it can be met (Gevers 1986; Moulin 1990d). Alternatively, we could imagine each agent in turn controlling an equal share of the social endowment and the technology, obtaining the *equal-division free-access upper bound* (Moulin 1990d; Yoshihara 1998). This definition can be generalized by imagining each group of agents in turn controlling a proportion of the social endowment equal to its relative size in the economy and the technology (Foley 1967). This yields the *equal-division free-access core*. There are economies with a concave production function in which no allocation is envy-free, efficient, and meets the equal-division free-access upper bound (Moulin 1990c). However, the bound is met on that domain by selections from the Pareto solution, in particular by the constant-returns-to-scale-equivalent solution defined later (Mas-Colell 1980; Moulin 1987b). For nowhere-decreasing returns-to-scale, the equal-division free-access bound becomes a lower bound: here, no sub-solution of the Pareto solution satisfies no envy for trades and meets the bound

(Moulin 1987b). Systematic studies of lower and upper bounds are Moulin (1990a, e, 1991, 1992b).

For one-input one-output production economies, *an allocation is proportional* if there are prices such that each agent, facing these prices, maximizes his preferences at his component of the allocation. These allocations can be used to define another lower bound on welfares (Maniquet 1996b, 2002; see also Roemer and Silvestre 1993).

The *constant-returns-to-scale lower bound* is defined, for each agent, by reference to the best bundle he could achieve if given access to a constant-returns-to-scale technology, the same for all agents; the *work-alone lower bound* is defined for each agent, by reference to the best bundle he could obtain if given access to the actual technology but under the obligation to provide bundles to the other agents to which he would not prefer his own (Fleurbaey and Maniquet 1996a, 1999).

Another study relating bounds in a class of two-good economies with convex production sets, the identical preferences lower bound and the free-access upper bound is due to Watts (1999).

Equal Opportunities as Equal, or Equivalent, Choice Sets

The notion of 'equal opportunities' is of course central in the theory of economic justice (for a general discussion, see Fleurbaey 1995c). The expression has been given a variety of meanings. In economies affected by uncertainty, it may mean 'equal treatment *ex ante*'. Uncertainty may also be endogenously generated by an allocation rule. Consider the problem of allocating an indivisible good. A lottery giving all agents equal chances might be deemed equitable *ex ante* although the final allocation may well appear inequitable. Alternatively, if agents' opportunities today are determined by decisions they made yesterday, equal opportunities may mean that they all had access to the same set of decisions. It is often argued that, because of incentive considerations, we should not attempt to equalize end results but instead should limit ourselves to giving people equal chances to develop their potential. If we do so, equal opportunities are provided by the mechanism that converts the choices agents make into a final outcome.

Another way to give substance to the idea of equal opportunities is to let each agent choose his consumption bundle from a common choice set (for example, see Kolm 1973). For the list of choices they make to constitute a feasible allocation, one should have access to a 'rich enough' family of choice sets. In addition, one would prefer efficiency to hold whenever feasibility does. Let $\mathscr{B}$ be a family of choice sets. An allocation is an *equal-opportunity allocation relative to* $\mathscr{B}$ (Thomson 1994a) if there is a member of $\mathscr{B}$ on which each agent maximizes his preferences at his component of the allocation. Such an allocation is of course envy-free. *The family* $\mathscr{B}$ *is satisfactory* on a domain if the resulting equal-opportunity allocations are always efficient. Under standard assumptions on preferences, the *equal-income Walrasian family* is satisfactory.

Another concept, *equal-opportunity–equivalence relative to a family* $\mathscr{B}$ *of choice sets*, generalizes the reasoning underlying egalitarian-equivalence. Check, whether, for some member of $\mathscr{B}$, each agent is indifferent between what he receives and the bundle he prefers in that set (Thomson 1994a). For the family of linear choice sets, and adding efficiency, we obtain any efficient allocation such that each agent finds his component of it indifferent to the best bundle he could achieve if endowed with an equal share of the social endowment and given access to a constant-returns-to-scale technology, the same for all agents (Mas-Colell 1980). Such an allocation is a *constant returns-to-scale equivalent allocation*. Other solutions are obtained by having all agents face a hypothetical technology obtained from the actual one by imagining the productivity of one specific factor of production (alternatively, of some subset of the factors of production) to be multiplied by some number, or by introducing a fixed cost of some factor of production (alternatively, introducing a fixed cost proportional to some fixed vector). Radial expansions and contractions of the production set can also be considered. An application of the concept is by Nicolò and Perea (2005).

The next definition generalizes proposals by Archibald and Donaldson (1979) and Varian (1976). An allocation *exhibits no envy of opportunities relative to a family $\mathscr{B}$ of choice sets* if for each agent, there is a member of $\mathscr{B}$ that contains the agent's maximizer on the union of everyone's sets (Thomson 1994a). For the family of linear choice sets, the resulting solution coincides with the equal-income Walrasian solution. If $\mathscr{B}$ is the family of $|N|$-lists of bundles, we obtain a concept that generalizes both no envy and egalitarian-equivalence. An allocation is *envy-free– equivalent* if there is a list of reference bundles, one for each agent, such that each agent is indifferent between his component of the allocation and his reference bundle and he finds his reference bundle at least as desirable as anyone else's reference bundle (Pazner 1977).

Monotonicity

Monotonicity properties are quite strong when imposed in conjunction with no envy and even no domination. Indeed, (*a*) no selection from the no-domination and Pareto solution is *resource-monotonic* (Moulin and Thomson 1988); (*b*) no selection from the no envy and Pareto solution is *population-monotonic* (Kim 2004); (*c*) no selection from the no domination and Pareto solution satisfies *welfare-domination under preference-replacement* (Thomson 1996). Other versions of these results are available, some of which involving significantly weaker distributional requirements (Geanakoplos and Nalebuff 1988; Moulin and Thomson 1988; Maniquet and Sprumont 2000; Kim 2001). However, if preferences satisfy gross substitutability and all goods are normal, the equal-division Walrasian solution is an example of a selection from the no envy and Pareto solution that is *resource-monotonic* and *population-monotonic* (Moulin and Thomson 1988; Fleurbaey 1995c).

On the other hand, no special assumptions are required for the existence of selections from the egalitarian-equivalence and Pareto solution that are *resource-monotonic*, or *population monotonic*, or satisfy *welfare-domination under preference-replacement*. Other rules based on the notion of equal-opportunity equivalence have these properties as well (Thomson 1987).

For economies with quasi-linear preferences satisfying certain additional assumptions, the Shapley value can provide the basis for a solution that is *resource-monotonic* (Moulin 1992a). The Shapley value has in fact proved useful on other domains to obtain this and other desirable properties of rules, although at the price of no envy, egalitarian-equivalence, and their variants.

The solidarity requirement can be applied to the joint replacement of resources and preferences (Sprumont 1996).

Technology-monotonicity is satisfied by certain selections from the egalitarian-equivalence and Pareto solution. For two goods, a characterization of a particular one is obtained by imposing it together with a few other minimal requirements. Suppose first that good 1 is used to produce good 2 according to a nowhere-decreasing-returns-to-scale technology. Given a group N of agents with preferences defined on $\mathbb{R}^2_+$, given some social endowment of good 1, which can be consumed as such or used as input in the production of good 2, the *equal-division free-access lower bound solution* selects the set of allocations such that each agent finds his bundle at least as desirable as the best bundle he could achieve if endowed with an equal share of the social endowment and given access to the technology.

Under alternative assumptions on technologies, (*a*) the only selection from the equal-division free-access lower bound and Pareto solution satisfying *Pareto-indifference* and *technology-monotonicity* is the constant-returns-to-scale–equivalence solution; (*b*) parallel characterizations hold for selections from the equal-division free-access upper bound (Moulin 1987b, 1990d).

Although in (*a*), the bounds on welfares are individual bounds, the solution that emerges happens to satisfy the requirement that no group of agents should be able to make each of its members at least as well off, and at least one of them better off, if each of its members is endowed with an equal share of the social endowment and the group is given access to the technology. A similar strengthening holds for (*b*).

F

Suppose now that resources and technologies both change. Dutta and Vohra (1993) require of a solution that if the set of feasible profiles of welfare levels enlarges, each allocation chosen initially should be welfare-dominated by some allocation chosen after the change, and that each allocation chosen after the change should welfare dominate some allocation chosen initially. Let us refer to this requirement as *opportunity-monotonicity*. The requirement of *r-equity* is that in an exchange economy in which there is only some amount of good *r* to divide, equal division should be chosen. Dutta and Vohra consider an invariance requirement that also depends on the choice of a good, say *r*, so we call it *r-invariance*. It is not motivated by normative considerations, so we only note that it is a weak version of an invariance requirement that has been important in the theory of implementation. The results are: up to Pareto-indifference, (*a*) the *r*-egalitarian equivalence and Pareto solution is the only selection from the Pareto solution satisfying *r-equity* and *opportunity-monotonicity*; (*b*) on the sub-domain of exchange economies, it is the only selection from the Pareto solution satisfying *r-equity*, *r-invariance* and *opportunity-monotonicity*.

Economies with production. In situations where agents are differentiated by their input contributions, a first monotonicity requirement is that if the contribution of an agent increases, he should end up at least as well off as he was initially. In situations in which agents differ in their productivities, a corresponding requirement is that if an agent's productivity increases, then again, he should end up at least as well off as he was initially.

The solidarity requirement, applied to the *joint* replacement of preferences and population in conjunction with the self-explanatory *replication-invariance*, leads to the selection from the egalitarian-equivalence solution for which the reference bundle is proportional to the social endowment (Sprumont and Zhou 1999; these authors also prove a version of this result for a model with infinitely many agents modelled as a continuum).

Economies with individual endowments. If the issue is that of allocating gains from trade, an appealing requirement is that when an agent's endowment increases, he should end up at least as well off as he was initially, *endowment monotonicity*. Another is that under the same hypotheses, nobody else should be made worse off than he was initially, *no negative effects on others*.

It is easy to define selections from the individual-endowments lower-bound and Pareto solution that are *own-endowment monotonic*. However, there are impossibilities too: (*a*) no selection from the no envy in trades and Pareto solution satisfies either *endowment monotonicity* or *no negative effect on others* (Thomson 1987); (*b*) no selection from the egalitarian-equivalence and Pareto solution satisfies *no negative effect on others* (Thomson 1987).

The appropriate expression of *population-monotonicity* here is that the welfares of all agents who are present before and after the change should be affected in the same direction. The Walrasian solution violates the property. However, the selections from the egalitarian-equivalence in trades and Pareto solution obtained by requiring the reference trade to lie on a monotone path satisfy the requirement. They also meet the individual-endowments lower bound (Thomson 1995a).

Consistency and Related Properties

Here, we also consider situations in which both the population of agents and the resources may vary, but this time, our focus is on a variety of invariance properties. These properties can be interpreted as formalizing trade-offs between equity and efficiency objectives with objectives of informational simplicity.

A converse of *replication-invariance*, *division-invariance*, says that if an allocation that is chosen for a replica economy happens to be a replica allocation (of the same order), then the model allocation should be chosen for the model economy.

The central notion, *consistency*, was defined in Section 1. Conversely, given some allocation that is feasible for some economy, check whether the restriction of the allocation to each subgroup of

two agents is chosen for the problem of allocating between them what they have received in total. If the answer is always yes, then one can say that each agent is treated fairly in relation to each other agent; then, *converse consistency* requires that the allocation itself should be chosen for the initial economy.

The Pareto solution is *consistent*. If preferences are smooth and corners excluded, it is also *conversely consistent* (Goldman and Starr 1982). The no-envy solution is both *consistent* and *conversely consistent*. The egalitarian-equivalence solution is *consistent* but not *conversely consistent*. This is also true for the equal-division Walrasian solution although, if preferences are smooth and corners excluded, this solution is *conversely consistent*.

We have the following characterizations: (*a*) if a sub-solution of the equal-division core is *replication-invariant*, then it is a sub-solution of the equal-division Walrasian solution (this is because under replication, the core 'shrinks' to the set of Walrasian allocations; Debreu and Scarf 1963; Thomson 1988; Nagahisa 1994, gives full characterizations of the Walrasian solution); (*b*) if a sub-solution of the group no-envy solution is *replication-invariant*, then it is a sub-solution of the equal-division Walrasian solution (Varian 1974); (*c*) under smoothness, if a sub-solution of the equal-division lower bound and Pareto solution is *replication-invariant* and *consistent*, then it is a sub-solution of the equal-division Walrasian solution (Thomson 1988); (*d*) under smoothness, if a sub-solution of the equal-division lower bound and Pareto solution is *anonymous* and *conversely consistent*, then on the sub-domain of two-agent economies, it is a sub-solution of the equal-division Walrasian solution; if in fact coincidence occurs on that sub-domain, then it contains the equal-division Walrasian solution for all other cardinalities (Thomson 1995b).

Consistency has been studied in economies with a large number of agents modelled as a continuum (Zhou 1992). For economies with possibly satiated preferences, a characterization of the 'equal-slack Walrasian solution' (Mas-Colell 1992) is available (Thomson and Zhou 1993). This solution differs from the standard Walrasian notion in that each agent's income is the sum of the value of his endowment at the prices announced by the auctioneer (they may have negative or 0 components) and a supplement, the same for all agents, which, like prices, is determined endogenously. Economies with both atoms and an atomless sector have also been studied (Zhou 1992; Shitovitz 1992).

Juxtaposition-invariance says that if an efficient allocation happens to be obtained by juxtaposing two allocations that are chosen for two sub-economies with equal per-capita social endowments, then it should be chosen (Thomson 1988). Under smoothness of preferences, the equal-division Walrasian solution is the only sub-solution of the Pareto solution satisfying a weak symmetry property, *juxtaposition-invariance*, and *consistency* (Maniquet 1996a).

In formulating *consistency* for a production economy, the issue arises of how to define the opportunities open to a group of agents after the members of the complementary group leave with their bundles. The simplest idea is to translate the production set by the sum of the bundles the departing agents took with them. Standard classes of technologies are not closed under this operation however, and adjustments have to be made to ensure that the 'reduced' production set is admissible. For economies with one-input one-output and inelastic demands, characterizations of *proportional cost sharing* and *serial cost sharing* (which can be understood as an extension of the Shapley value) are available (Moulin and Shenker 1994).

The *equal-wage-equivalent and Pareto solution* selects the efficient allocations for which there is a reference wage such that each agent finds his bundle indifferent to the best bundle he could achieve by maximizing his preferences on a budget set defined by this wage. The *output-egalitarian-equivalence and Pareto solution* selects the efficient allocations that each agent finds indifferent to a common consumption consisting of only some amount of the output.

Under appropriate assumptions on technologies, (*a*) the former is the only *essentially single-valued* selection from the constant-returns-to-scale lower bound solution satisfying *Pareto*

F

indifference, *equal welfares for equal preferences* (self-explanatory), *contraction independence* (as in bargaining theory), and *consistency*; (*b*) the latter is the only *essentially single-valued* selection from the work-alone lower bound solution satisfying *Pareto indifference*, *equal welfares for equal preferences*, and *consistency* (Fleurbaey and Maniquet 1999).

Roemer (1986a, 1986b, 1988) formulates consistency requirements with respect to changes in the number of goods.

When a solution is not *consistent*, it has a *minimal consistent enlargement* (Thomson 1994d). For instance, the minimal *consistent* enlargement of the equal-division lower bound and Pareto solution is 'essentially' the Pareto solution. That of the Ω-egalitarian-equivalence and Pareto solution is 'essentially' the egalitarian-equivalence and Pareto solution. The *maximal consistent sub-solution* of a solution can be defined in a symmetric way provided the solution contains at least one *consistent* solution.

Notions of consistency have been proposed for economies with individual endowments (Thomson 1992; van den Nouweland et al. 1996; Serrano and Volij 1998; Korthues 2000).

Indivisible Goods

Estate or divorce settlements often involve items that cannot be divided (houses, family heirlooms), or can only be divided at a cost that would make the division undesirable (silverware). Other examples are positions in schools or organs for transplant patients. We call such goods 'objects'. We consider situations in which the social endowment also contains some amount of an infinitely divisible good, 'money'. We focus on situations in which each agent can consume at most one object. An illustration is the problem of allocating rooms to students in the house they share, and specifying how much each of them should contribute to the rent.

Let A be a set of objects. Each agent has preferences defined over $\mathbb{R} \times A$ (or over $\mathbb{R}_+ \times A$). They are continuous and strictly monotonic with respect to money, and such that the switch from any object to any other object can be compensated by an appropriate adjustment in the consumption of money. The simplest case is when there are as many objects as agents. Situations where there are fewer objects than agents are accommodated by introducing a 'null object'; there are always enough copies of the null object for each agent to end up with one (real or null) object. If there are fewer agents than objects, some objects are unassigned. In some applications, it is natural to require that the null object should not be assigned until all real objects are, even if these objects are undesirable, or undesirable for some agents. They could be tasks to be assigned to housemates that none of them enjoys performing; alternatively, some of them may find a given task enjoyable and the others not (cooking).

Punctual requirements. It is clear that if consumptions of money have to be non-negative, envy-free allocations may not exist. Otherwise, or if the social endowment of money is large enough, existence holds (Svensson 1983a; Maskin 1987; Alkan et al. 1991; Tadenuma and Thomson 1993; Ichiishi and Idzik 1999; Su 1999). For quasi-linear preferences, several algorithms leading to envy-free allocations are available (Aragones 1995; Klijn 2000; Ünver 2003; Abdulkadiroğlu et al. 2004). Remarkably, envy-free allocations are always efficient (Svensson 1983b). A variety of selections from the no-envy solution have been proposed (Tadenuma 1989, 1994; Alkan et al. 1991; Aragones 1995; Tadenuma and Thomson 1995).

Egalitarian-equivalent and efficient allocations exist very generally, when preferences are defined over $\mathbb{R} \times A$ and the compensation assumption holds. When preferences are defined over $\mathbb{R}_+ \times A$, existence holds under similar assumptions as the ones guaranteeing that of envy-free allocations. Just as in the classical case, there are economies in which all egalitarian-equivalent and efficient allocations violate no-envy.

The case of one object is special, and the solution that selects the envy-free allocation at which the winner receives the least amount of money has a number of interesting properties and has been characterized on the basis of these properties. This allocation is egalitarian-equivalent, with

the losers' bundle serving as reference bundle (Tadenuma and Thomson 1993; Thomson 1998).

The *Walrasian solution* can easily be adapted to the present model but here, an allocation is an equal-income Walrasian allocations if and only if it is envy-free and efficient, and if and only if it is group envy-free (Svensson 1983b).

Another requirement is that each agent should be made at least as well off as he would be at the (essentially) unique envy-free allocation of the hypothetical economy in which everyone had his preferences. If $|N| = 2$, meeting this *identical-preferences lower bound* is equivalent to no-envy, but if $|N| > 2$, the identical-preferences lower bound is weaker (Bevia 1996a). Thus, this concept gives us another chance of obtaining positive results when no-envy is too demanding. Unfortunately, there are quasi-linear economies with equal numbers of objects and agents in which all egalitarian-equivalent and efficient allocations violate not only no-envy, as already noted, but in fact the identical-preferences lower bound. When there are more objects than agents, an allocation may be envy-free and efficient without meeting the identical-preferences lower bound, but it does meet the variant of the lower bound obtained by using only the objects that are assigned. No-envy remains incompatible with this bound however (Thomson 2003).

Relational requirements. Selections from the no-envy solution exist that satisfy a form of *money-monotonicity* (Alkan et al. 1991). Any selection from the egalitarian-equivalence and Pareto solution obtained by fixing the reference object is *money-monotonic*.

Object-monotonicity, the requirement that when additional objects become available, all agents should end up at least as well off as they were initially, makes sense if the objects are desirable or when there are undesirable objects, they do not have to be assigned. To study it, in specifying an economy, we now have to allow the numbers of objects and agents to differ. Then, an envy-free allocation is not necessarily efficient and we explicitly impose efficiency. Unfortunately, no selection from the no-envy and Pareto solution is *object-monotonic*, even if preferences are quasi-linear (Alkan 1994).

Suppose now that all real objects have to be assigned before any null object is, independently of whether they are desirable. For instance, objects may be activities that some agents enjoy and others do not, but these activities have to be carried out if there are enough agents for that, an example mentioned earlier. Even if preferences are quasi-linear, no selection from a natural weakening of the identical-preferences lower bound and Pareto solution is *weakly object monotonic*, that is, such that the welfares of all agents should be affected in the same direction by an enlargement of the set of objects (Thomson 2003).

Even if preferences are quasi-linear, no selection from the no-envy solution satisfies *welfare-domination under preference-replacement* (Thomson 1998).

A first requirement in the context of a variable population is that if the social endowment of money is non-negative and the objects are all desirable, none of the agents initially present should benefit from the arrival of additional agents. Even if preferences are quasi-linear, *population-monotonicity* is incompatible with no-envy (Alkan 1994; Moulin 1990b). In fact, an agent could be better off at any envy-free allocation than if he were alone, so that the *free-access upper bound* is incompatible with no-envy.

If there is a single object, which is desirable, and the social endowment of money is zero, a *population-monotonic* selection from the identical-preferences lower bound and Pareto solution can be defined, based on the Shapley value (Moulin 1990b; Bevia 1996c). Other positive results can be obtained for that case.

The selection from the egalitarian-equivalence and Pareto solution obtained by requiring the reference bundle to contain a fixed object is *weakly population-monotonic* (the arrival of new agents affects the welfares of all existing agents in the same direction), but it is not guaranteed to be a selection from the no-envy solution any more. In fact, no selection from the no-envy solution is *weakly population-monotonic* (Tadenuma and Thomson 1995). Weaker requirements pertaining to changes in resources or population are defined and investigated by Alkan (1994).

Turning to *consistency*, we have the following result: if a sub-solution of the no-envy solution is *neutral* (that is, invariant under exchanges of bundles that leave all agents indifferent) and *consistent*, then in fact, it coincides with the no-envy solution (Tadenuma and Thomson 1991). As always, the no-envy solution is *conversely consistent*, but many proper sub-solutions of it are too (as well as *neutral*). On the other hand, the Pareto solution is not (unless the objects are identical). However, if a sub-solution of the no-envy solution is *neutral*, *bilaterally consistent*, and *conversely consistent*, then in fact it coincides with the no-envy solution (Tadenuma and Thomson 1991).

The identical-preferences lower bound solution is *conversely consistent* but not *consistent*. The *minimal consistent enlargement* of its intersection with the Pareto solution is the Pareto solution itself. This is true when there is at most one object, when there are multiple identical objects, and when there are multiple and possibly different objects. The maximal *consistent* sub-solution of the identical-preferences lower bound and Pareto solution is the no-envy solution (Bevia 1996a).

Related models. When each agent can consume several objects (in addition to the infinitely divisible good), the situation is quite different from what it is in the one-object-per-agent case, unless severe additional restrictions are imposed on preferences. In fact, many of the special relations that exist in the one-object-per-agent case disappear, and the situation resembles the classical situation (Tadenuma 1996; Haake et al. 2002).

For preferences that have additive representations, a rule proposed by Knaster (1946) is generalized by Steinhaus (1949) and advocated by Samuelson (1980). An alternative is the selection from the egalitarian-equivalence and Pareto solution obtained by choosing the null object as reference object. Interestingly, it is a selection from the no-envy solution (Willson 2003). Each is *money-monotonic* and satisfies a form of *object-monotonicity*.

Even if preferences are quasi-linear and no other fairness requirement is imposed, no selection from the Pareto solution is *population-monotonic* (Bevia 1996b). In contrast to the one-object-per-person case, there are *consistent* sub-solutions of the no-envy and Pareto solution, and *converse consistency* becomes a much stronger requirement. Characterizations have been obtained under an additional invariance requirement on solutions (Bevia 1998). The *population-monotonicity* of rules that select lotteries is examined by Ehlers and Klaus (2003b).

When monetary compensations are not possible. This situation has recently been much studied, mainly in the one-object-per-agent case when preferences are strict. It is clear that punctual requirements of fairness such as no-envy and egalitarian-equivalence are not generally achievable here (think of situations where all agents have the same preferences). However, most of our relational requirements remain meaningful. The main lesson of the literature is that they can be satisfied, but in a rather limited way, by sequential priority rules and variants. If the objective is to respect an exogenously given priority order of agents, then of course more positive results can be obtained (Svensson 1994a; Balinski and Sönmez 1999; Ergin 2000, 2002; Ehlers and Klaus 2006, 2007; Kesten 2006).

Now, imagine that agents can consume several objects. Herreiner and Puppe (2002) propose a maximin-type criterion, and define an iterative procedure that produces, among the efficient allocations, the one that is best according to this criterion (see also Ramaekers 2006). In that situation, no selection from the Pareto solution satisfies *welfare-domination under preference replacement* (Klaus and Miyagawa 2001).

Brams and Fishburn (2000) for $|N| = 2$ and Edelman and Fishburn (2001) for $|N| > 2$ examine the special case when agents have the same preferences over individual objects but possibly different preferences over sets of objects. Brams et al. (2003) drop the assumption that preferences over individual objects are the same, and propose, in addition to requirements related to no-envy, some that are based on comparing the numbers of objects received by the various agents.

The possibility that agents are endowed with objects is considered by Shapley and Scarf (1974), and situations when some objects are initially individually owned and others are commonly owned (residential housing on a university campus being an illustrative example; kidney exchange is

another application) are discussed by Roth et al. (2004) and Sönmez and Ünver (2005).

Various notions of efficiency for rules that select lotteries are examined by Hylland and Zeckhauser (1979), Demko and Hill (1988), Abdulkadiroğlu and Sönmez (1998), and Bogomolnaia and Moulin (2001, 2002).

When objects cannot be transferred. Consider the problem of allocating a single infinitely divisible good, 'money', among agents characterized by variables that cannot be transferred (talent or handicaps for examples), and thus can be thought of 'objects'. How should money be divided to compensate agents for possible differences in these variables? This question, formulated by Fleurbaey (1994, 1995a), has given rise to a large literature. For a detailed survey, see Fleurbaey and Maniquet (2008).

Single-Peaked Preferences

Consider the problem of allocating a social endowment of an infinitely divisible and non-disposable commodity among a group of agents whose preferences over $\mathbb{R}_+$ are single-peaked: up to some critical level, his *peak amount*, an increase in an agent's consumption increases his welfare but beyond that level, the opposite holds. Since there is no possibility of disposal, the social endowment has to be fully distributed. If the sum of the peak amounts is greater than the social endowment, 'there is not enough', and for efficiency, no agent should consume more than his peak amount. If the inequality goes the other way, 'there is too much'; here, for efficiency, no agent should consume less than his peak amount (Sprumont 1991).

Punctual requirements. Efficient allocations meeting the equal-division lower bound, or no-envy, in fact both, always exist. The equal-division core and the group-no-envy solution may be empty, but natural variants of these solutions are not.

A number of interesting rules can be defined: the commodity can be divided proportionally to the peak amounts, or so that all agents' consumptions are at the same distance from their peak amounts subject to non-negativity, or so that the sizes of their upper contour sets at their assigned consumptions are equal, or as equal as possible. The following rule, called the *uniform rule*, will be central: if there is not enough, and given $\lambda \geq 0$, assign to each agent the amount he prefers in $[0,\lambda]$; choose λ so that the sum of these assignments is equal to the social endowment; if there is too much, given $\lambda \geq 0$, assign to each agent the amount he prefers in $[\lambda,\infty]$; here too, choose λ so that the sum of these assignments is equal to the social endowment.

The uniform rule depends only on the profile of peak amounts – it satisfies the *peak-only* requirement – and it is the only subsolution of the no-envy and Pareto solution to do so (Thomson 1994c). Also, it is the only selection from the Pareto solution minimizing (*a*) the difference between the smallest amount anyone receives and the greatest amount anyone receives; (*b*) alternatively, the variance of the amounts they all receive (Schummer and Thomson 1997).

Relational requirements. Here, the natural expression of the idea of solidarity when the social endowment varies is that all agents should be made at least as well off as they were initially *or* that they should all be made at most as well off. This requirement is incompatible with no-envy (or with the equal-division lower bound). This is because a change in the social endowment can be so disruptive that it turns an economy in which there is not enough to one in which there is too much, or converse. This suggests limiting its application to situations in which no such switches occur, yielding *one-sided resource-monotonicity*. This property is much less demanding. Solidarity requirements with respect to changes in population or preferences can similarly be modified by limiting their application to situations in which the direction of the inequality between the sum of the peak amounts and the social endowment is not reversed by the change under consideration. We add the suffix 'one-sided' to indicate the weaker versions so defined. We also consider *separability*, which says that given two economies having a group of agents in common, if the agents in this group receive the same aggregate amount in both, then each of them should receive the same amount in both.

We have the following characterizations, some of which require that each preference relation be such that if its peak amount is positive, there is an amount greater than the peak amount that is indifferent to 0. The uniform rule is (*a*) the only selection from the no-envy and Pareto solution to be *one-sided resource-monotonic* (Thomson 1994b); (*b*) the only selection from the no-envy and Pareto solution to satisfy *replication-invariance* and *one-sided welfare-domination under preference-replacement* (Thomson 1997); (*c*) the only selection from the no-envy and Pareto solution to be *replication-invariant* and *one-sided population-monotonic* (Thomson 1995a); (*d*) the only selection from the no-envy and Pareto solution to be *resource-continuous* and *separable* (Chun 2003, 2006). (*d*) the smallest (in terms of inclusion) subsolution of the no-envy and Pareto solution to be *resource upper hemi-continuous* and *consistent* (Thomson 1994c); (*e*) the only *single-valued* selection from the equal-division lower bound and Pareto solution to be *replication-invariant* and *consistent*, or to be *anonymous* and *conversely consistent* (Thomson 1995c).

Many refinements and variants of these results are available (Sönmez 1994; Klaus 1997, 1999, 2006; Dagan 1996; Moulin 1999; Herrero and Villar 1998, 2000; Ehlers 2002a, b; Kesten 2004b). An application to a pollution problem is by Kıbrıs (2003).

Related models. Fairness issues have been analysed for the variant of the model obtained by introducing individual endowments (Thomson 1995c; Klaus 1997, 2001; Klaus et al. 1997; Moreno 2002).

For economies with both individual endowments and a social endowment, different ways of adapting the punctual fairness requirements have been proposed, and issues of *monotonicity*, with respect to the individual endowments and the social endowment, in addition to *consistency* and *population-monotonicity*, have been addressed (Thomson 1995c; Klaus 1997; Herrero 2002). In these studies, a rule that is the natural extension of the uniform rule most frequently emerges.

A multi-commodity version of the single-peaked assumption is easily defined. For such a model, a generalization of the equal-slacks Walrasian solution (Mas-Colell 1992) is axiomatized along the lines of Schummer and Thomson's (1997) axiomatization of the uniform rule (Amoros 1999). A probabilistic version of the uniform rule is characterized by Sasaki (1997).

Non-Homogeneous Continuum

Here, we consider the problem of dividing a heterogeneous commodity modelled as measure space, each agent having preferences defined over the measurable subsets, and the question being how to select partitions consisting of measurable subsets, one for each agent. Think of a cake on which frosting and decorations are distributed unevenly. Often, this commodity is embedded in a finite-dimensional Euclidean space: an example is land.

Punctual requirements. In such situations, equal division has no economic meaning, even when it can be defined in physical terms (surface area, say, or weight). However, our central criteria (no-envy; egalitarian-equivalence) remain applicable. A large literature concerns preferences that can be represented by atomless measures, a somewhat restrictive assumption that precludes complementarities between different parts of the dividend. Additional topological and geometric criteria are sometimes meaningful (Hill 1983). The construction of iterative procedures leading to partitions satisfying some fairness requirement, exactly or in some approximate sense, has been important in the literature, but until recently, efficiency had often been ignored.

If no restrictions are imposed on preferences apart from continuity and monotonicity with respect to set inclusion, envy-free and efficient partitions may not exist (Berliant et al. 1992). However, when preferences are representable by atomless measures, they do (Weller 1985). An existence result for group envy-free partitions is also available (Berliant et al. 1992).

An interesting special case is the one-dimensional case when the dividend is an interval that has to be partitioned into subintervals, one for each agent. It has many applications: division of an interval of time, a length of road, and so

on. When preferences are represented by atomless measures, envy-free partitions exist (Woodall 1980), but in fact existence then holds under much weaker assumptions (Stromquist 1980; Su 1999). In the case of a closed curve, the situation is much less satisfactory (Barbanel and Brams 2005; Thomson 2007). Under monotonicity of preferences, no-envy implies efficiency (Berliant et al. 1992).

Under continuity and strict monotonicity of preferences, egalitarian-equivalent and efficient allocations exist (Berliant et al. 1992).

When preferences are represented by atomless measures, the $\frac{1}{n}$ – *lower-bound* is that for each agent, the value to him of his assignment should be at least $\frac{1}{n}$ times his value of the dividend. Some of the early literature searched for partitions such that for each agent, this bound is met as an equality. Given a list $\alpha \in \Delta^N$ of 'shares', the α – *lower-bound* is that for each $i \in N$, the value to agent i of his assignment should be at least α_i times his value of the dividend. Partitions satisfying these notions and generalizations exist (Berliant et al. 1992; Barbanel and Zwicker 2001; Reijnierse and Potters 1998). An existence result is available when preferences are representable by atomless concave capacities (Maccheroni and Marinacci 2003). The existence of envy-free partitions is also known for a more general notion of a partition, where agents receive 'fractional' consumptions of each point of the dividend (Akin 1995).

A succession of attempts at generalizing to more than two agents the classical two-person divide-and-choose scheme (one agent divides and the other chooses one of the two pieces; the divider receives the other), have been made over the years that generate partitions that are either envy-free or meet the $\frac{1}{n}$ – *lower-bound*. It took many years until an algorithm that produces an envy-free partition in the n-person case, for arbitrary n, was discovered (Brams and Taylor 1995). None of the solutions proposed necessarily attains efficiency.

Brams and Taylor (1996) survey the literature. Robertson and Webb (1998) focus on algorithms and pay little attention to efficiency. On the other hand, Barbanel (2005) provides an in-depth analysis of the shape of the image of the set of feasible partitions in a Euclidean space of dimension equal to the number of agents, using their measures as representations of their preferences. It offers characterizations of its subset of efficient points. It also gives existence results for efficient and envy-free partitions.

Other Domains

We conclude this survey by tying it to literatures concerning other models but also addressing fairness issues. They concern (*a*) the Arrovian model of extended sympathy; (*b*) rights assignments; (*c*) quasi-linear social choice; (*d*) intertemporal allocation; (*e*) public choice from an interval or a closed curve when agents have single-peaked preferences; (*f*) public good production; (*g*) cost sharing; (*h*) queuing, scheduling, and sequencing; (*i*) matching.

See Also

- ▶ Efficient Allocation
- ▶ Equality of Opportunity
- ▶ Justice
- ▶ Justice (New Perspectives)
- ▶ Shapley Value

Bibliography

Abdulkadiroğlu, A., and T. Sönmez. 1998. Random serial dictatorship and the core from random endowments in house allocation problems. *Econometrica* 66: 689–701.

Abdulkadiroğlu, A., and T. Sönmez. 1999. House allocation with existing tenants. *Journal of Economic Theory* 88: 233–260.

Abdulkadiroğlu, A., T. Sönmez, and U. Ünver. 2004. Room assignment-rent division: A market approach. *Social Choice and Welfare* 22: 515–538.

Akin, E. 1995. Vilfredo Pareto cuts the cake. *Journal of Mathematical Economics* 24: 23–44.

Alkan, A. 1994. Monotonicity and envyfree assignments. *Economic Theory* 4: 605–616.

Alkan, A., G. Demange, and D. Gale. 1991. Fair allocation of indivisible goods and criteria of justice. *Econometrica* 59: 1023–1039.

The author would like to thank the National Science Foundation for its support under grant SES 0214691.

Amoros, P. 1999. Efficiency and income redistribution in the single-peaked preference model with several commodities. *Economics Letters* 63: 341–349.

Aragones, E. 1995. A derivation of the money Rawlsian solution. *Social Choice and Welfare* 12: 267–276.

Archibald, P., and D. Donaldson. 1979. Notes on economic inequality. *Journal of Public Economics* 12: 205–214.

Aumann, R., and B. Peleg. 1974. A note on Gale's example. *Journal of Mathematical Economics* 1: 209–211.

Balinski, M., and T. Sönmez. 1999. A tale of two mechanisms: Student placement. *Journal of Economic Theory* 84: 73–94.

Barbanel, J. 2005. *The geometry of efficient Fair division*. Cambridge: Cambridge University Press.

Barbanel, J., and S. Brams. 2005. *Cutting a pie is not a piece of cake*. Mimeo: New York University.

Barbanel, J., and W. Zwicker. 2001. Two applications of a theorem of Dvoretsky, Wald, and Wolfovitz to cake division. *Theory and Decision* 43: 639–650.

Berliant, M., K. Dunz, and W. Thomson. 1992. On the fair division of a heterogeneous commodity. *Journal of Mathematical Economics* 21: 201–216.

Bevia, C. 1996a. Identical preferences lower bound and consistency in economies with indivisible goods. *Social Choice and Welfare* 13: 113–126.

Bevia, C. 1996b. Population monotonicity in a general model with indivisible goods. *Economics Letters* 50: 91–97.

Bevia, C. 1996c. Population monotonicity in economies with one indivisible good. *Mathematical Social Sciences* 32: 125–137.

Bevia, C. 1998. Fair allocation in a general model with indivisible goods. *Review of Economic Design* 3: 195–213.

Bogomolnaia, A., and H. Moulin. 2001. A new solution to the random assignment problem. *Journal of Economic Theory* 100: 295–328.

Bogomolnaia, A., and H. Moulin. 2002. A simple random assignment problem with a unique solution. *Economic Theory* 19: 298–317.

Brams, S., and P. Fishburn. 2000. Fair division of indivisible items between two people with identical preferences: Envy-freeness, Pareto-optimality, and equity. *Social Choice and Welfare* 17: 247–267.

Brams, S., and A.D. Taylor. 1995. An envy-free cake division protocol. *American Mathematical Monthly* 102: 9–18.

Brams, S., and A.D. Taylor. 1996. *Fair Division: From Cake-Cutting to Dispute Resolution*. Cambridge: Cambridge University Press.

Brams, S., P. Edelman, and P. Fishburn. 2003. Fair division of indivisible items. *Theory and Decision* 55: 147–180.

Champsaur, P., and G. Laroque. 1981. Fair allocations in large economies. *Journal of Economic Theory* 25: 269–282.

Chaudhuri, A. 1985. Formal properties of interpersonal envy. *Theory and Decision* 18: 301–312.

Chaudhuri, A. 1986. Some implications of an intensity measure of envy. *Social Choice and Welfare* 3: 255–270.

Chichilnisky, G., and W. Thomson. 1987. The Walrasian mechanism from equal division is not monotonic with respect to variations in the number of consumers. *Journal of Public Economics* 32: 119–124.

Chun, Y. 2003. One-sided population-monotonicity, separability, and the uniform rule. *Economics Letters* 78: 343–349.

Chun, Y. 2006. The separability principle in economies with single-peaked preferences. *Social Choice and Welfare* 26: 239–253.

Chun, Y., and W. Thomson. 1988. Monotonicity properties of bargaining solutions when applied to economics. *Mathematical Social Sciences* 15: 11–27.

Dagan, N. 1995. *Consistent solutions in exchange economies: A characterization of the price mechanism*. Mimeo: Pompeu Fabra University.

Dagan, N. 1996. A note on Thomson's characterization of the uniform rule. *Journal of Economic Theory* 69: 255–261.

Daniel, T. 1975. A revised concept of distributional equity. *Journal of Economic Theory* 11: 94–109.

Daniel, T. 1978. Pitfalls in the theory of fairness: Comment. *Journal of Economic Theory* 19: 561–564.

Debreu, G., and H. Scarf. 1963. A limit theorem on the core of an economy. *International Economic Review* 4: 235–246.

Demko, S., and T. Hill. 1988. Equitable distribution of indivisible objects. *Mathematical Social Sciences* 16: 145–158.

Diamantaras, D. 1992. On equity with public goods. *Social Choice and Welfare* 9: 141–157.

Diamantaras, D., and W. Thomson. 1990. An extension and refinement of the no-envy concept. *Economics Letters* 33: 217–222.

Dutta, B., and R. Vohra. 1993. A characterization of egalitarian equivalence. *Economic Theory* 4: 465–479.

Edelman, P., and P. Fishburn. 2001. Fair division of indivisible items among people with similar preferences. *Mathematical Social Sciences* 41: 327–347.

Ehlers, L. 2002a. Resource-monotonic allocation when preferences are single-peaked. *Economic Theory* 20: 113–131.

Ehlers, L. 2002b. *A characterization of the uniform rule without Pareto-optimality.* Mimeo: University of Montreal.

Ehlers, L., and B. Klaus. 2003a. Coalitional strategy-proofness and resource-monotonicity for multiple assignment problems. *Social Choice and Welfare* 21: 265–280.

Ehlers, L., and B. Klaus. 2003b. Probabilistic assignments of identical indivisible objects and the probabilistic uniform correspondence. *Review of Economic Design* 8: 249–268.

Ehlers, L., and B. Klaus. 2003c. Resource-monotonicity for house allocation problems. *International Journal of Game Theory* 32: 545–560.

Ehlers, L., and B. Klaus. 2006. Efficient priority rules. *Games and Economic Behavior* 55: 372–384.

Ehlers, L., and B. Klaus. 2007. Consistent house allocation. *Economic Theory* 30: 561–574.
Ehlers, L., B. Klaus, and S. Pápai. 2002. Strategy-proofness and population-monotonicity for house allocation problems. *Journal of Mathematical Economics* 38: 329–339.
Ergin, H. 2000. Consistency in house allocation problems. *Journal of Mathematical Economics* 34: 77–97.
Ergin, H. 2002. Efficient resource allocation on the basis of priorities. *Econometrica* 70: 2489–2497.
Feldman, A., and A. Kirman. 1974. Fairness and envy. *American Economic Review* 64: 995–1005.
Fleurbaey, M. 1994. On fair compensation. *Theory and Decision* 36: 277–307.
Fleurbaey, M. 1995a. The requisites of equal opportunity. In *Social choice, welfare, and ethics*, ed. W.A. Barnett, H. Moulin, M. Salles, and N. Schofield. Cambridge: Cambridge University Press.
Fleurbaey, M. 1995b. Three solutions for the compensation problem. *Journal of Economic Theory* 65: 505–521.
Fleurbaey, M. 1995c. Equal opportunity or equal social outcomes. *Economics and Philosophy* 11: 25–55.
Fleurbaey, M., and F. Maniquet. 1996a. Cooperative production: A comparison of welfare bounds. *Games and Economic Behavior* 17: 200–208.
Fleurbaey, M., and F. Maniquet. 1996b. Fair allocation with unequal production skills: The no-envy approach to compensation. *Mathematical Social Sciences* 32: 71–93.
Fleurbaey, M., and F. Maniquet. 1999. Cooperative production with unequal skills: The solidarity approach to compensation. *Social Choice and Welfare* 16: 569–583.
Fleurbaey, M., and F. Maniquet. 2008. Compensation and responsibility. In *Handbook of social choice and welfare*, vol. 2, ed. K. Arrow, A. Sen, and K. Suzumura. Amsterdam: North-Holland.
Foley, D. 1967. Resource allocation and the public sector. *Yale Economic Essays* 7: 45–98.
Geanakoplos, J., and B. Nalebuff. 1988. *On a fundamental conflict between equity and efficiency*. Mimeo: Yale University.
Gevers, L. 1986. Walrasian social choice: Some simple axiomatic approaches. In *Social choice and public decision making, essays in honor of K.J. Arrow*, ed. W.P. Heller, R.M. Starr, and D.A. Starrett. Cambridge: Cambridge University Press.
Ginés, M., and F. Marhuenda. 1996. Cost monotonic mechanisms. *Investigaciones Economicas* 20: 89–103.
Goldman, S., and R. Starr. 1982. Pairwise, *t*-wise and Pareto-optimality. *Econometrica* 50: 593–606.
Goldman, S., and C. Sussangkarn. 1980. On equity and efficiency. *Economics Letters* 5: 29–31.
Green, J. 1972. On the inequitable nature of core allocations. *Journal of Economic Theory* 4: 132–143.
Haake, C.-J., M. Raith, and F. Su. 2002. Bidding for envy-freeness: A procedural approach to *n*-player fair-division problems. *Social Choice and Welfare* 19: 723–749.
Herreiner, D., and C. Puppe. 2002. A simple procedure for finding equitable allocations of indivisible goods. *Social Choice and Welfare* 19: 415–430.
Herrero, C. 2002. General allocation problems with single-peaked preferences: Path-independence and related topics. *Spanish Economic Review* 4: 19–40.
Herrero, C., and A. Villar. 1998. The equal-distance rule in allocation problems with single-peaked preferences. In *Current trends in economic theory and applications*, ed. R. Aliprantis and N. Yannelis. Berlin: Springer.
Herrero, C., and A. Villar. 2000. An alternative characterization of the equal-distance rule for allocation problems with single-peaked preferences. *Economics Letters* 66: 311–317.
Hill, T.P. 1983. Determining a fair border. *American Mathematical Monthly* 90: 438–442.
Hylland, A., and R. Zeckhauser. 1979. The efficient allocations of individual to positions. *Journal of Political Economy* 87: 293–314.
Ichiishi, T., and A. Idzik. 1999. Equitable allocation of divisible goods. *Journal of Mathematical Economics* 32: 389–400.
Jaskold-Gabszewicz, J.-J. 1975. Coalitional fairness of allocations in pure exchange. *Econometrica* 43: 661–668.
Kesten, O. 2004a. *Coalition strategy-proofness and resource monotonicity for house allocation problems*. Mimeo: Carnegie-Melon University.
Kesten, O. 2004b. *More on the uniform rule: Characterizations without Pareto-optimality*. Mimeo: Carnegie-Melon University.
Kesten, O. 2006. On two competing mechanisms for priority-based allocation problems. *Journal of Economic Theory* 27: 155–171.
Khan, A., and H. Polemarchakis. 1978. Unequal treatment in the core. *Econometrica* 46: 1475–1481.
Kıbrıs, O. 2003. Permit allocation problems. *Social Choice and Welfare* 20: 353–362.
Kim, H. 2001. Essays on fair allocations, Ph.D. thesis.
Kim, H. 2004. Population monotonicity for fair allocation problems. *Social Choice and Welfare* 23: 59–70.
Klaus, B. 1997. The characterization of the uniform reallocation rule without side-payments. In *Game theoretic applications to economics and operations research*, ed. T. Parthasarathy, B. Dutta, J.A.M. Potters, T.E.S. Raghavan, D. Ray, and A. Sen. Dordrecht: Kluwer-Academic.
Klaus, B. 1999. *The role of replication-invariance: Two answers concerning the problem of fair division when preferences are single-peaked*. Mimeo: University of Maastricht.
Klaus, B. 2001. Uniform allocation and reallocation revisited. *Review of Economic Design* 6: 85–98.
Klaus, B. 2006. A note on the separability principle in economies with single-peaked preferences. *Social Choice and Welfare* 26: 255–261.
Klaus, B., and E. Miyagawa. 2001. Strategy-proofness, solidarity, and consistency for multiple assignment problems. *International Journal of Game Theory* 30: 421–435.

Klaus, B., H. Peters, and T. Storcken. 1997. Reallocation of an infinitely divisible good. *Economic Theory* 10: 305–333.

Kleinberg, N. 1980. Fair allocations and equal incomes. *Journal of Economic Theory* 23: 189–200.

Klijn, F. 2000. An algorithm for envy-free allocations in an economy with indivisible objects and money. *Social Choice and Welfare* 17: 201–215.

Knaster, B. 1946. Sur le problème du partage pragmatique de H. Steinhaus. *Annales de la Société Polonaise de Mathématique* 19: 228–230.

Kolm, S. 1972. *Justice et Equité*. Paris: Editions du Centre National de la Recherche Scientifique, 1972. English edition. Cambridge, MA: MIT Press, 1988.

Kolm, S. 1973. Super-equité. *Kyklos* 26: 841–843.

Kolpin, V. 1991a. Resolving open questions on the λ^*-envy-free criterion. *Economics Letters* 36: 17–20.

Kolpin, V. 1991b. Equity and the core. *Mathematical Social Sciences* 22: 137–150.

Korthues, B. 2000. Characterization of an extended Walrasian concept for open economies. *Journal of Mathematical Economics* 33: 449–461.

Maccheroni, F., and M. Marinacci. 2003. How to cut a pizza fairly: Fair division with decreasing marginal evaluations. *Social Choice and Welfare* 20: 457–465.

Maniquet, F. 1996a. Horizontal equity and stability when the number of agents is variable in the fair division problem. *Economics Letters* 50: 85–90.

Maniquet, F. 1996b. Allocation rules for a commonly owned technology: The average cost lower bound. *Journal of Economic Theory* 69: 490–508.

Maniquet, F. 1999. A strong incompatibility between efficiency and equity in non-convex economies. *Journal of Mathematical Economics* 32: 467–474.

Maniquet, F. 2002. A study of proportionality and simplicity in the cooperative production problem. *Review of Economic Design* 7: 1–15.

Maniquet, F., and Y. Sprumont. 2000. On resource-monotonicity in the fair division problem. *Economics Letters* 68: 299–302.

Mas-Colell, A. 1980. Remarks on the game theoretic analysis of a simple distribution of surplus problems. *International Journal of Game Theory* 9: 125–140.

Mas-Colell, A. 1987. On the second welfare theorem for anonymous net trades ain exchange economies with many agents. In *Information, incentives and economic mechanisms*, ed. T. Groves, R. Radner, and S. Reiter. Minneapolis: University of Minnesota Press.

Mas-Colell, A. 1992. Equilibrium theory with possibly satiated preferences. In *Equilibrium and dynamics; essays in honor of D. Gale*, ed. M. Majumdar. London: Macmillan.

Maskin, E. 1987. On the fair allocation of indivisible goods. In *Arrow and the foundations of the theory of economic policy*, ed. G. Feiwel. London: Macmillan.

Moreno, B. 2002. Single-peaked preferences, population-monotonicity and endowments. *Economics Letters* 75: 87–95.

Moulin, H. 1987a. A core selection for pricing a single-output monopoly. *Rand Journal of Economics* 18: 397–407.

Moulin, H. 1987b. The pure compensation problem: Egalitarianism versus laissez-fairism. *Quarterly Journal of Economics* 101: 769–783.

Moulin, H. 1990a. Uniform preference externalities: Two axioms for fair allocation. *Journal of Public Economics* 43: 305–326.

Moulin, H. 1990b. Monotonic surplus-sharing and the utilization of common property resources. In *Game theory and applications*, ed. T. Ichiishi, A. Neyman, and Y. Tauman. New York: Academic Press.

Moulin, H. 1990c. Fair division under joint ownership: Recent results and open problems. *Social Choice and Welfare* 7: 149–170.

Moulin, H. 1990d. Joint ownership of a convex technology: Comparisons of three solutions. *Review of Economic Studies* 57: 439–452.

Moulin, H. 1990e. Interpreting common ownership. *Recherches Economiques de Louvain* 56: 303–326.

Moulin, H. 1991. Welfare bounds in the fair division problem. *Journal of Economic Theory* 54: 321–337.

Moulin, H. 1992a. An application of the Shapley value to fair division with money. *Econometrica* 60: 1331–1349.

Moulin, H. 1992b. All sorry to disagree: A general principle for the provision of non-rival goods. *Scandinavian Journal of Economics* 94: 37–51.

Moulin, H. 1993. On the fair and coalitionproof allocation of private goods. In *Frontiers of game theory*, ed. K. Binmore, A. Kirman, and P. Tani. Cambridge, MA: MIT Press.

Moulin, H. 1995. *Cooperative microeconomics: A game-theoretic introduction*. Princeton: Princeton University Press.

Moulin, H. 1996. Cost sharing under increasing returns: A comparison of simple mechanisms. *Journal of Economic Theory* 13: 225–251.

Moulin, H. 1999. Rationing a commodity along fixed paths. *Journal of Economic Theory* 84: 41–72.

Moulin, H. 2003. *Fair division and collective welfare*. Cambridge, MA: MIT Press.

Moulin, H., and J. Roemer. 1989. Public ownership of the external world and private ownership of self. *Journal of Political Economy* 97: 347–367.

Moulin, H., and S. Shenker. 1994. Average cost pricing versus serial cost sharing: An axiomatic comparison. *Journal of Economic Theory* 64: 178–207.

Moulin, H., and W. Thomson. 1988. Can everyone benefit from growth? Two difficulties. *Journal of Mathematical Economics* 17: 339–345.

Nagahisa, R. 1994. A necessary and sufficient condition for Walrasian social choice. *Journal of Economic Theory* 62: 186–208.

Nicolò, A., and A. Perea. 2005. Monotonicity and equal-opportunity equivalence in bargaining. *Mathematical Social Sciences* 49: 221–243.

Otsuki, M. 1980. On distribution according to labor: A concept of fairness in production. *Review of Economic Studies* 47: 945–958.

Pazner, E. 1977. Pitfalls in the theory of fairness. *Journal of Economic Theory* 14: 458–466.

Pazner, E., and D. Schmeidler. 1974. A difficulty in the concept of fairness. *Review of Economic Studies* 41: 441–443.

Pazner, E., and D. Schmeidler. 1978. Egalitarian equivalent allocations: A new concept of economic equity. *Quarterly Journal of Economics* 92: 671–687.

Piketty, T. 1994. Existence of fair allocations in economies with production. *Journal of Public Economics* 55: 391–405.

Ramaekers, E. 2006. *Fair allocation of indivisible goods without monetary compensations*. Mimeo: University of Namur.

Reijnierse, J.H., and J.A.M. Potters. 1998. On finding an envy-free Pareto-optimal division. *Mathematical Programming* 83: 291–311.

Robertson, J., and W. Webb. 1998. *Cake-cutting algorithms*. Natick: A.K. Peters.

Roemer, J. 1986a. The mismarriage of bargaining theory and distributive justice. *Ethics* 97: 88–110.

Roemer, J. 1986b. Equality of resources implies equality of welfare. *Quarterly Journal of Economics* 101: 751–784.

Roemer, J. 1988. Axiomatic bargaining on economic environments. *Journal of Economic Theory* 45: 1–35.

Roemer, J. 1996. *Theories of distributive justice*. Cambridge, MA: Harvard University Press.

Roemer, J., and J. Silvestre. 1993. The proportional solution for economies with both private and public ownership. *Journal of Economic Theory* 59: 426–444.

Roth, A., T. Sönmez, and U. Ünver. 2004. Kidney exchange. *Quarterly Journal of Economics* 119: 457–488.

Samuelson, W. 1980. The object distribution problem revisited. *Quarterly Journal of Economics* 94: 85–98.

Sasaki, H. 1997. Randomized uniform allocation mechanism and single-peaked preferences of indivisible good. Working Paper. Waseda University.

Schmeidler, D., and K. Vind. 1972. Fair net trades. *Econometrica* 40: 637–642.

Schummer, J., and W. Thomson. 1997. Two derivations of the uniform rule. *Economics Letters* 55: 333–337.

Serrano, R., and O. Volij. 1998. Axiomatizations of neoclassical concepts for economies. *Journal of Mathematical Economics* 30: 87–108.

Shapley, L., and H. Scarf. 1974. On cores and indivisibility. *Journal of Mathematical Economics* 1: 23–37.

Shitovitz, B. 1992. Coalitional fair allocations in smooth mixed markets with an atomless sector. *Mathematical Social Sciences* 25: 27–40.

Sönmez, T. 1994. Consistency, monotonicity and the uniform rule. *Economics Letters* 46: 229–235.

Sönmez, T., and U. Ünver. 2005. House allocation with existing tenants: An equivalence. *Games and Economic Behavior* 52: 153–185.

Sprumont, Y. 1991. The division problem with single-peaked preferences: A characterization of the uniform allocation rule. *Econometrica* 59: 509–519.

Sprumont, Y. 1996. Axiomatizing ordinal welfare egalitarianism when preferences vary. *Journal of Economic Theory* 68: 77–110.

Sprumont, Y., and L. Zhou. 1999. Pazner-Schmeidler rules in large society. *Journal of Mathematical Economics* 31: 321–339.

Steinhaus, H. 1949. Sur la division pragmatique. *Econometrica* 17: 315–319.

Stromquist, N. 1980. How to cut a cake fairly. *American Mathematics Monthly* 87: 640–644.

Su, F. 1999. Rental harmony: Sperner's lemma in fair division. *American Mathematical Monthly* 106: 930–942.

Svensson, L.-G. 1983a. On the existence of fair allocations. *Zeitschrift für National Oekonomie* 43: 301–308.

Svensson, L.-G. 1983b. Large indivisibilities: An analysis with respect to price equilibrium and fairness. *Econometrica* 51: 939–954.

Svensson, L.-G. 1994a. Queue allocation of indivisible goods. *Social Choice and Welfare* 11: 323–330.

Svensson, L.-G. 1994b. σ-optimality and fairness. *International Economic Review* 35: 527–531.

Tadenuma, K. 1989. *On the single-valuedness of a solution for the problem of fair allocation in economies with indivisibilities*. Mimeo: University of Rochester.

Tadenuma, K. 1994. *On strongly envy-free allocations in economies with indivisible goods*. Mimeo: University of Rochester.

Tadenuma, K. 1996. Trade-off between equity and efficiency in a general model with indivisible goods. *Social Choice and Welfare* 13: 445–450.

Tadenuma, K., and W. Thomson. 1991. No-envy and consistency in economies with indivisible goods. *Econometrica* 59: 1755–1767.

Tadenuma, K., and W. Thomson. 1993. The fair allocation of an indivisible good when monetary compensations are possible. *Mathematical Social Sciences* 25: 117–132.

Tadenuma, K., and W. Thomson. 1995. Refinements of the no-envy solution in economies with indivisible goods. *Theory and Decision* 39: 189–206.

Thomson, W. 1978. *Monotonic allocation mechanisms: Preliminary results*. Mimeo: University of Rochester.

Thomson, W. 1983a. Equity in exchange economies. *Journal of Economic Theory* 29: 217–244.

Thomson, W. 1983b. The fair division of a fixed supply among a growing population. *Mathematics of Operations Research* 8: 319–326.

Thomson, W. 1987. Monotonic allocation rules. Revised 1989. University of Rochester.

Thomson, W. 1988. A study of choice correspondences in economies with a variable number of agents. *Journal of Economic Theory* 46: 237–254.

Thomson, W. 1992. *Consistency in exchange economies*. Mimeo: University of Rochester.

Thomson, W. 1994a. Notions of equal, and equivalent, opportunities. *Social Choice and Welfare* 11: 137–156.

Thomson, W. 1994b. Resource-monotonic solutions to the problem of fair division when preferences are single-peaked. *Social Choice and Welfare* 11: 205–223.
Thomson, W. 1994c. Consistent solutions to the problem of fair division when preferences are single-peaked. *Journal of Economic Theory* 63: 219–245.
Thomson, W. 1994d. Consistent extensions. *Mathematical Social Sciences* 28: 35–49.
Thomson, W. 1995a. Population-monotonic solutions to the problem of fair division when preferences are single-peaked. *Economic Theory* 5: 229–246.
Thomson, W. 1995b. *The theory of fair allocation*. Mimeo: University of Rochester.
Thomson, W. 1995c. *Endowment monotonicity in economies with single-peaked preferences*. Mimeo: University of Rochester.
Thomson, W. 1996. *The replacement principle in classical economies with private goods*. Mimeo: University of Rochester.
Thomson, W. 1997. The replacement principle in private good economies with single-peaked preferences. *Journal of Economic Theory* 76: 145–168.
Thomson, W. 1998. The replacement principle in economies with indivisible goods. *Social Choice and Welfare* 15: 57–66.
Thomson, W. 1999. Welfare-domination and preference-replacement: A survey and open questions. *Social Choice and Welfare* 16: 373–394.
Thomson, W. 2003. On monotonicity in economies with indivisible goods. *Social Choice and Welfare* 21: 195–205.
Thomson, W. 2006a. *Population-monotonic allocation rules*. Mimeo: University of Rochester.
Thomson, W. 2006b. *Consistent allocation rules*. Mimeo: University of Rochester.
Thomson, W. 2006c. *Fair allocation rules*. Mimeo: University of Rochester.
Thomson, W. 2007. Children crying at birthday parties; why? *Economic Theory* 31: 501–521.
Thomson, W., and L. Zhou. 1993. Consistent allocation rules in atomless economies. *Econometrica* 61: 575–587.
Ünver, U. 2003. *Market mechanisms for fair division with indivisible objects and money.* Mimeo: University of Pittsburg.
van den Nouweland, A., B. Peleg, and S. Tijs. 1996. Axiomatic characterizations of the Walras correspondence for generalized economies. *Journal of Mathematical Economics* 25: 355–372.
Varian, H. 1974. Equity, envy, and efficiency. *Journal of Economic Theory* 9: 63–91.
Varian, H. 1975. Distributive justice, welfare economics, and the theory of fairness. *Philosophy and Public Affairs* 4: 223–247.
Varian, H. 1976. Two problems in the theory of fairness. *Journal of Public Economics* 5: 249–260.
Vohra, R. 1992. Equity and efficiency in non-convex economies. *Social Choice and Welfare* 9: 185–202.
Watts, A. 1999. Cooperative production: A comparison of lower and upper bounds. *Journal of Mathematical Economics* 32: 317–331.
Weller, D. 1985. Fair division of measurable space. *Journal of Mathematical Economics* 14: 5–17.
Willson, S. 2003. Money-egalitarian-equivalent and gain-maximin allocations of indivisible items with monetary compensation. *Social Choice and Welfare* 20: 247–259.
Woodall, J.R. 1980. Dividing a cake fairly. *Journal of Mathematical Analysis and Applications* 78: 233–247.
Yannelis, N. 1983. Existence and fairness of value allocation without convex preferences. *Journal of Economic Theory* 31: 283–292.
Yoshihara, N. 1998. Characterizations of public and private ownership solutions. *Mathematical Social Sciences* 35: 165–184.
Young, P. 1994. *Equity.* Princeton: Princeton University Press.
Zhou, L. 1992. Strictly fair allocations and Walrasian equilibria in large exchange economies. *Journal of Economic Theory* 57: 158–175.

Fair Division

Vincent P. Crawford

The theory of fair division is concerned with the design of procedures for allocating a bundle of goods among n persons who are perceived to have equal rights to the goods. Both equity (according to criteria discussed below) and efficiency are sought. The theory is of interest primarily because its approach to allocation problems enjoys some important advantages over the alternative approach suggested by neoclassical welfare economics, and because studying the sense in which procedures actually in use are equitable is a good way to learn about popular notions of equity.

The modern theory of fair division has it origins in papers by Steinhaus (1948) and Dubins and Spanier (1961), who described methods (attributed by Steinhaus in part to S. Banach and K. Knaster) for sharing a perfectly divisible 'cake' among n people. In the method described by Steinhaus, the people are ordered (randomly, if desired) and the first person cuts a slice from the cake. Then each other person, in turn, may

diminish the slice if he wishes. The last person to diminish the slice must take it as his share, with the slice reverting to the first person if no one chooses to diminish it. The process then continues, sharing the remainder of the cake in the same way among those people who have not yet received a share.

In the closely related method described by Dubins and Spanier, one person passes a knife continuously over the cake, at each instant determining a well-defined slice, which grows over time. The first other person to indicate his willingness to accept the slice then determined by the knife's location receives it as his share. The process then continues as before.

These *n*-person fair-division schemes are in the spirit of the classical two-person method of divide and choose, in which one person divides the cake into two portions and the other then chooses between them. Neither *n*-person scheme, however, is a true generalization of the two-person method. Steinhaus (1950) proposed a three-person scheme (formalized and generalized to *n* persons by Kuhn 1967) that is a true generalization. In this scheme, one person divides the cake into *n* portions and the others announce which of the portions are acceptable to them. Then, if it is possible to give each of the others a share acceptable to him, this is done. Otherwise, it is possible to assign a share to the divider in such a way that it is still feasible to give each other person 1/*n*th of the cake in his own estimation. This share is assigned, and the process then continues as before.

Each of these schemes is fair in the sense that, under reasonably general conditions (see Kuhn 1967), it allows each person to ensure, independent of the others' behaviour, that he will obtain at least 1/*n*th of the total value of the cake in his estimation. In the Steinhaus (1948) method, if a person is called upon to cut, he takes a slice with 1/*n*th the value of the original cake; and a person given an opportunity to diminish a slice reduces it to 1/*n*th value, if possible, or does nothing if it already has value 1/*n*th or less. In the method described by Dubins and Spanier, each person indicates his willingness to accept any slice whose value reaches 1/*n*th of the total value of the cake. Finally, in the method of Steinhaus (1950), the divider divides the cake into *n* portions, each acceptable to him, and the others declare acceptable all portions they deem to have at least 1/*n*th of the value of the entire cake.

These results are of considerable interest, but are incomplete in several ways. First, they ignore the question of efficiency, which is central to the problem of designing allocation mechanisms.

Second, although it does not involve interpersonal comparisons, the notion of fairness employed is inherently cardinal, and therefore difficult to make operational. This obscures a major advantage of the fair-division approach over that of neoclassical welfare economics.

Finally, when operationally meaningful notions of fairness are employed in an environment with nontrivial efficiency issues, the fact that each person has a strategy that ensures him at least his share of the cake does not guarantee that allocations resulting from strategic behaviour are fair: a person might give up the social desideratum of fairness to get more of the goods he desires.

The modern theory of fair division answers these criticisms by studying the implications of rational behaviour and employing a different concept of equity. A fair procedure is defined as one that always yields a fair allocation, in the sense formalized by Foley (1967): an allocation is *fair* if and only if no person prefers any other person's share to his own.

Kolm (1972) and Crawford (1977) (see also Luce and Raiffa 1957, and Crawford and Heller 1979) use this notion to formalize the sense in which the two-person method of divide and choose is fair. They characterize the perfect-equilibrium strategies when the divider (D) knows the preferences of the chooser (C) and show that in equilibrium, D divides so that he is indifferent about C's choice and C then chooses as D would prefer. The resulting allocation is fair, in Foley's sense, but not generally efficient unless D and C have identical preferences. The allocation is, however, efficient in the set of fair allocations.

These results establish an operationally meaningful sense in which the two-person divide-and-choose method is fair, and show that it has some

tendency toward efficiency. However, when preferences are common knowledge, the role of divider is an advantage, in the sense that the divider always weakly prefers his allocation to what he would receive if he were chooser. This follows from the facts that the game always yields a fair allocation and the divider can divide so that any desired fair allocation is the result. Further, *n*-person versions of the divide-and-choose method need not even yield fair allocations.

Crawford (1979) and Crawford (1980) study schemes that improve upon the classical divide-and-choose method while preserving its good points. In the two-person scheme studied in Crawford (1980), D offers C a choice between a proposal of D's choosing and equal division, instead of making him choose between a proposal and its complement. The resulting perfect-equilibrium outcomes are both fair and efficient, under reasonable assumptions; the role of divider is still an advantage, but less so than in the classical divide-and-choose method. These results extend, in part, to the *n*-person case.

In the *n*-person scheme studied in Crawford (1979), the role of divider in the scheme of Crawford (1980) is auctioned off. This completely eliminates the asymmetry of roles, and yields perfect-equilibrium allocations that are both efficient and egalitarian-equivalent, in the sense of Pazner and Schmeidler (1978): an allocation is *egalitarian-equivalent* if and only if it is indifferent, for all people, to equal division of some (not necessarily feasible) bundle of goods. However, although egalitarian-equivalence shares many of fairness's advantages as an equity notion, egalitarian-equivalent allocations need not be fair.

Despite their flaws, the schemes just described share several advantages over the traditional approach of choosing an allocation that maximizes a neoclassical social welfare function.

First, they deal with notions of equity that (like efficiency) do not involve interpersonal comparisons and have an objective meaning.

Second, their prescriptions are implementable in a stronger sense than those of neoclassical welfare economics. The classical welfare theorems establish that a competitive equilibrium is efficient and that, under reasonable assumptions, any efficient allocation can be obtained as a competitive equilibrium for suitably chosen initial endowments. But finding the endowments that yield the allocation that maximizes social welfare is informationally virtually equivalent to computing the entire optimal allocation. By contrast, the fair-division approach often allows the specification of procedures that are independent of the details of the environment but still yield equitable and efficient allocations.

Finally, most of the procedures studied in the literature on fair division are self-administered, in the sense that they can be implemented without a referee. This is difficult to formalize, but clearly important in practice.

See Also

- ▶ Envy
- ▶ Equality
- ▶ Fairness

Bibliography

Crawford, V. 1977. A game of fair division. *Review of Economic Studies* 44(2): 235–247.

Crawford, V. 1979. A procedure for generating Pareto-efficient egalitarian-equivalent allocations. *Econometrica* 47(1): 49–60.

Crawford, V. 1980. A self-administered solution of the bargaining problem. *Review of Economic Studies* 47(2): 385–392.

Crawford, V., and W. Heller. 1979. Fair division with indivisible commodities. *Journal of Economic Theory* 21(1): 10–27.

Dubins, L., and E. Spanier. 1961. How to cut a cake fairly. *American Mathematical Monthly* 68(1): 1–17.

Foley, D. 1967. Resource allocation and the public sector. *Yale Economic Essays* 7(1): 45–98.

Kolm, S. 1972. *Justice et équité*. Paris: Editions du Centre National de la Recherche Scientifique.

Kuhn, H. 1967. Chapter 2. On games of fair division. In *Essays in honor of Oskar Morgenstern*, ed. M. Shubik. Princeton: Princeton University Press.

Luce, R., and H. Raiffa. 1957. *Games and decisions: Introduction and critical survey*. New York: Wiley.

Pazner, E., and D. Schmeidler. 1978. Egalitarian equivalent allocations: A new concept of economic equity. *Quarterly Journal of Economics* 92(4): 671–687.

Steinhaus, H. 1948. The problem of fair division. *Econometrica* 16(1): 101–104.

Steinhaus, H. 1950. *Mathematical snapshots*. New York: Oxford University Press.

Fairness

Hal R. Varian

The issues of equity and efficiency are central aspects of most economic problems. In the political domain it often seems that concerns with equity – or at least distribution – often outweigh concerns with economic efficiency in discussion of policy alternatives. Despite this, most economic analysis has paid much more attention to issues of efficiency than to equity.

The notion of efficiency has been repeatedly refined in economics, and today the concept of a Pareto efficient allocation has a firm place in the economist's tool-kit. There is no similar agreement about the proper concept of 'equitable' or 'fair' allocations. This is not to say that proposals are lacking, and in this essay I will examine a few of the ideas concerning economic definitions of fairness and equity. Since I have provided a more detailed survey of contributions in this area elsewhere (Thomson and Varian 1985), I will focus more on the conceptual underpinnings, rather than the technical results.

Suppose that you had a bundle of goods to divide in a 'fair' way among n economic agents. How would you do it? In the absence of any further information, the natural choice is equal division. But even if equal division is a fair way to divide the bundle *initially*, it may not remain fair. If agents have different tastes, they will generally desire to trade the goods among themselves. Even though the initial allocation is symmetric, the final allocation will not necessarily inherit this desirable property of the original division.

What would be an economic definition of 'symmetry'? One proposal, due to Duncan Foley (1967), goes as follows: an agent i is said to *envy* another agent j if i prefers j's bundle to his own. An allocation in which no agent envies any other agent is known as an *envy-free* allocation. Equal division is, of course, envy-free, but there will typically be many other allocations that satisfy this symmetry property. Allocations that are both Pareto efficient and envy-free are particularly interesting since they are allocations that will not be disturbed by voluntary trade. An envy-free allocation is sometimes referred to as an 'equitable' allocation. An envy-free Pareto efficient allocation is often called a 'fair' allocation. The term 'envy-free' seems to me to be both more descriptive and less misleading.

But do Pareto efficient envy-free allocations necessarily exist? It is too much to ask for allocations that are both equitable and efficient? As it turns out, it is possible to show that a competitive equilibrium from equal division is necessarily an envy-free and efficient allocation. It is efficient by the First Theorem of Welfare Economics, and the envy-free property follows from the fact that equal division guarantees that all agents will have the same wealth.

Other sorts of allocative mechanisms may not necessarily preserve the symmetry of equal division. For example, it is easy to exhibit allocations in the core of an equal division market game in which some agent envies another. The particular feature of trade on a competitive market that is important is the fact that all agents have the same trading opportunities, and hence cannot in equilibrium prefer some other agent's choices to their own. This insight has been examined in detail by Schmeidler and Vind (1972) using the notion of 'fair net trades'.

The concept of envy-free allocations has been generalized in many different ways. For example, there is the idea of a 'coalitionally envy-free allocation', which requires that there is no *group* of agents that unanimously prefers some other group's bundle to their own. A closely related idea is that of an *egalitarian equivalent* allocation, which is one in which every agent is indifferent between the bundle he holds in that allocation and a bundle in some (hypothetical) equal division allocation.

There will typically exist envy-free Pareto efficient allocations that are not competitive equilibria with equal wealths, but an equal-wealth allocation turns out to be especially interesting in a number of ways. For example, only in an equal-wealth allocation does each agent have the same budget set, and thus have *equal trading opportunities*. Furthermore, it can be shown that when preferences vary continuously across the

population, the *only* Pareto efficient envy-free allocations are those with equal wealth.

The concept of envy-free allocations seems to work quite well as a formalization of the concept of symmetry when agents are themselves more or less symmetrically situated. However, when the agents are not themselves symmetric, the envy-free concept becomes somewhat forced. Consider, for example, the case of agents with severe handicaps. Do they not deserve some kind of special compensation for these handicaps in a 'fair' allocation? Shouldn't a diabetic's demand for insulin take precedence over a gourmet's demand for truffles?

These questions arise naturally when we consider models of production. For in this case agents with different abilities are like agents with different degrees of being handicapped. As Ronald Dworkin (1981) puts it: 'someone who cannot play basketball like Wilt Chamberlain... suffers from an (especially common) handicap.' How does the concept of an envy-free allocation generalize to production economies? First we should consider what we mean by stating that one agent envies another agent in a production context. Since one agent cannot directly consume another agent's leisure, the extension of the concept to production is not immediate. More formally, if one agent's consumption set is not identical with another's, the concept of envy-free allocation is not necessarily well defined.

The natural thing to do here is to consider what would happen if agents swapped not only consumption bundles but also labour commitments – in order to envy another agent, you not only have to desire his consumption, but you also have to be willing to work as much as he does.

But this definition has a serious problem which was first discovered by Pazner and Schmeidler (1974): it may be that there are no Pareto efficient envy-free allocations by this definition. The problem is that just because one agent is willing to work as much as another doesn't mean that he will be able to produce as much output as the other. When abilities are different, the concept of 'envy' needs some refinement.

One suggestion, made by Varian (1974), is to have agents compare their consumption-output bundles, not their consumption-input bundles. Thus in order to 'envy' another agent, I must be willing to produce as much as he produces, or, more generally, I have to produce output with the same value that he produces. This sort of envy comparison, happily, is consistent with Pareto efficiency. Another suggestion, due to Pazner and Schmeidler (1978), is that we consider allocations in which each agent has a consumption-leisure bundle that has equal value at the efficiency prices. Again, it can be shown that such allocations will always exist. In some sense these two proposals are at opposite extremes: Varian's suggestion favours the able, while Pazner and Schmeidler's favours the unable. Is there a natural intermediate concept that is in some sense more balanced? The answer is not known.

An area that is closely related to that of envy-free allocations is that of *games of fair division*. Everyone is familiar with the classic scheme of 'I divide and you choose' as a solution to two person division games. But what do you do if you want to divide a good (or a bundle of goods) among more than two agents? There have been several schemes proposed; Kuhn (1967) provides a nice survey of the early literature. Since this survey, there have been some further study of games of fair division and an increasing interest in the implementability of some of the equity concepts described above.

See Also

- ▶ Equity
- ▶ Fair division
- ▶ Justice
- ▶ Welfare economics

Bibliography

Dworkin, R. 1981. What is equality? Part 2: Equality of resources. *Philosophy and Public Affairs* 10: 283–345.

Foley, D. 1967. Resource allocation and the public sector. *Yale Economic Essays* 7: 45–98.

Kuhn, H. 1967. On games of fair division. In *Essays in honor of Oskar Morgenstern*, ed. M. Shubik. Princeton: Princeton University Press.

Pazner, E., and D. Schmeidler. 1974. A difficulty in the concept of fairness. *Review of Economic Studies* 41: 441–443.
Pazner, E., and D. Schmeidler. 1978. Decentralization and income distribution in socialist economies. *Economic Inquiry* 16(2): 257–264.
Schmeidler, D., and K. Vind. 1972. Fair net trades. *Econometrica* 40: 637–647.
Thomson, W., and H. Varian. 1985. Theories of justice based on symmetry. In *Social goals and social organization: Essays in honor of Elisha Pazner*, ed. Leo Hurwicz et al. New York: Oxford University Press.
Varian, H. 1974. Equity, envy, and efficiency. *Journal of Economic Theory* 9: 63–91.

Fall of AIG

William K. Sjostrom

Abstract

This article provides a brief overview of AIG's operations and explains why AIG suddenly collapsed. It then details the terms of the initial US government bailout and later restructurings. Finally, the article describes the regulatory gap exploited by AIG and ensuing regulatory reform.

Keywords

AIG; American International Group; CDO; CDS; Collateralized debt obligations; Credit default swaps; Derivatives; Tranching

JEL Classifications

K20; K22; G8; G28; G38

Introduction

In 2007, American International Group, Inc. (AIG), then the largest insurance company in the USA, generated $110 billion in total revenue and earned $8.9 billion in operating income. AIG ended 2007 with over $1 trillion in assets and $95.8 billion in shareholders equity. A mere nine and a half months later, however, AIG was on the verge of bankruptcy and had to be rescued by the US government through an $85 billion loan. Additional aid followed, and US government commitments ultimately grew to more than $180 billion.

This article provides a brief overview of AIG's operations and explains why AIG suddenly collapsed. It then details the terms of the initial US government bailout and later restructurings. Finally, the article describes the regulatory gap exploited by AIG and ensuing regulatory reform.

AIG's Pre-Bailout Operations

Overview

AIG is a holding company incorporated in Delaware, and its common stock is listed on the New York Stock Exchange. Prior to the bailout, AIG engaged, through its subsidiaries, in a broad range of insurance and insurance-related activities in the USA and abroad. AIG had operations in more than 130 countries, with about half of its revenues derived from its foreign operations. Its principal business units were General Insurance, Life Insurance & Retirement Services, Financial Services, and Asset Management. The General Insurance unit underwrote commercial property, casualty, workers' compensation, and mortgage guarantee insurance. The Life & Retirement Service unit provided individual and group life, payout annuities, endowment, and accident and health insurance policies. The Financial Services unit engaged in aircraft and equipment leasing, capital market transactions, consumer finance, and insurance premium finance. The Asset Management unit offered a wide variety of investment-related services and investment products to individuals, pension funds and institutions. AIG ranked tenth in the 2007 Fortune 500.

Table 1 summarizes AIG's operating performance by unit for the years ended 31 December 2005, 2006 and 2007.

AIG's Credit Default Swap Business

As Table 1 indicates, AIG's operating income dropped by $12.7 billion from 2006 to 2007

Fall of AIG, Table 1 AIG revenues and operating income, 2005–2007

(In millions)	2007	2006	2005
Revenues			
General insurance	$51,708	$49,206	$45,174
Life insurance and retirement	53,570	50,878	48,020
Financial services	(1,309)	7,777	10,677
Asset management	5,625	4,543	4,582
Other	457	483	344
Consolidation and eliminations	13	500	(16)
Total	$110,064	$113,387	$108,781
Operating income (loss)			
General insurance	$10,562	$10,412	$2,315
Life insurance and retirement	8,186	10,121	8,965
Financial services	(9,515)	383	4,424
Asset management	1,164	1,538	1,963
Other	(2,140)	(1,435)	(2,765)
Consolidation and eliminations	722	668	311
Total	$8,943	$21,687	$15,213

Source: AIG Annual Report (2007)

principally because of the $9.5 billion loss posted by its Financial Services unit. For the most part, this loss resulted from write-downs on the unit's credit default swap (CDS) business. AIG's CDS business was run by AIG subsidiaries AIG Financial Products Corp. and AIG Trading Group, Inc., and their respective subsidiaries (collectively, AIGFP) out of Connecticut and London. Because AIGFP's CDS business was at the heart of AIG's collapse, this section provides a short primer on CDSs and then describes the business.

A CDS is a privately negotiated contract where one party (the 'protection seller'), in exchange for a fee, agrees to compensate another party (the 'protection buyer') if a specified 'credit event' (such as bankruptcy or failure to pay) occurs with respect to a company (the 'reference entity') or debt obligation (the 'reference obligation'). CDSs have historically been transacted over-the-counter (OTC), meaning they were not traded on an exchange or cleared through a clearinghouse. They fall under the broader category of OTC derivatives, which includes interest rate, currency and commodities swaps.

CDSs are used for a variety of purposes, including hedging, speculation and arbitrage. For example, if a mutual fund wants to hedge its credit risk exposure on its $100 million of XYZ Inc. (XYZ) bonds that mature in five years, it can do so by entering into a five-year, $100 million CDS with a protection seller. The CDS would designate XYZ as the reference entity and XYZ's bonds as the reference obligation. It would define credit event as XYZ's bankruptcy or payment default on its bonds. In this example, the CDS would have a 'notional amount' of $100 million because that is the amount of protection provided by the CDS. In connection with writing the CDS, the protection seller would assess the likelihood of a credit event occurring during the next five years and set its fee for providing the protection accordingly. This fee is referred to as the CDS spread or premium and is expressed in basis points per annum on the notional amount of the CDS. The spread is typically payable quarterly. In this example, if the protection seller sets the spread at 100 basis points, the fund would pay the protection seller $250,000 per quarter during the five-year term of the CDS.

If no credit event occurs during the term of a CDS, the protection seller retains the premium payments and the parties go their separate ways. In this example, that means the protection seller would have grossed $5 million from writing the CDS ($250,000 per quarter multiplied by twenty quarters). If a credit event does occur during the CDS term, the protection seller is then obligated to compensate the protection buyer. Compensation occurs through either physical or cash settlement, depending on what the CDS specifies. If the CDS provides for physical settlement, it will specify types of 'deliverable obligations' that the protection seller is required to buy for par (full face value) upon delivery by the protection seller. In this example, assume the CDS provided for physical settlement and designated the XYZ bonds as the deliverable obligation. Following an XYZ credit event, the fund would transfer the $100 million face amount of XYZ bonds to the protection seller. The protection seller would then pay the fund $100 million, and the CDS would terminate. Obviously, XYZ bonds will have dropped in

value as a result of the credit event and, therefore, will be worth much less than par.

If the CDS provides for cash settlement, the parties agree on a market value for the reference obligation. The protection seller then pays the protection buyer the difference between the market value and the par value of the reference obligation. In this example, assume that the market value of the reference obligation dropped to 25% of par following the credit event. The protection seller would then pay the fund $75 million ($100 million par value less the $25 million market value) and the CDS would terminate.

A prominent risk inherent in a CDS faced by a protection buyer is counterparty credit risk. Counterparty credit risk is the risk that a protection seller will be unable or unwilling to make the payment due under a CDS following a credit event. To address counterparty credit risk, a CDS may require the protection seller to post collateral with the protection buyer equal to a specified percentage of the notional amount of the CDS. If the market price of the referenced obligation declines by a certain amount or the credit rating of the referenced obligation is downgraded, the CDS would typically require the protection seller to post additional collateral as these happenings generally indicate a perceived increase in the probability of a credit event occurring. The initial collateral percentage typically varies depending on the protection seller's credit rating. The higher its credit rating, the lower the collateral percentage. This is because a higher credit rating indicates higher credit quality and, therefore, a lower chance that a protection seller will default on its obligations under the CDS. The CDS will typically provide for an automatic increase in the collateral percentage for any downgrades to the protection seller's credit rating during the term of the CDS.

AIGFP's CDS business consisted largely of selling protection on 'super senior risk tranches of diversified pools of loans and debt securities' (AIG Annual Report 2007). Deciphering what exactly this means requires a basic understanding not only of CDSs but also of asset-backed securities. Asset-backed securities are securities backed by a discrete pool of financial assets such as commercial loans, residential mortgage loans, credit card receivables or student loans. Asset-backed securities are created through the process of securitization.

The most relevant type of asset-backed securities when it comes to AIG's collapse is residential mortgage-backed securities. The typical securitization process for these securities is as follows. It starts with a borrower applying to a lender (either directly or through a broker) for a mortgage loan to purchase a home or refinance an existing loan. Assuming the application is approved, the lender funds the loan as part of the purchase or refinancing closing. Then the lender sells the loan to an institution called an arranger (sometimes also called an issuer). The arranger then sells the loans – and oftentimes similar loans it has purchased from other lenders –to a newly formed special purpose vehicle (SPV). The SPV funds the purchase of the loans by selling investors debt obligations representing claims to the cash flows from the pool of residential mortgage loans owned by the SPV. These obligations are 'asset-backed securities' because they are 'backed' or supported by a financial asset (the mortgage loans). The SPV uses the cash flows from the pool of mortgage loans (primarily monthly loan payments) to service the debt it issued investors to buy the loans.

Often, the SPV divides the debt securities it issues into different tranches reflecting different levels of seniority or payment priority. For example, the SPV could issue three different classes of debt securities: a senior class, a mezzanine class and a junior class. The SPV's indenture (the document that specifies the terms of the debt securities) would then provide that obligations (interest and principal) owed to the senior class are to be paid first, followed by those owed to the mezzanine class, with the junior class to be paid last. If all amounts owed on the loans or other financial assets owned by the SPV are paid timely, the SPV will have sufficient funds to meet its obligations with respect to all three classes. If funds are insufficient, the junior class is the first not to get paid, followed by the mezzanine class. The senior class would only not get paid if the SPV's shortfall exceeds amounts owed to the junior class and the mezzanine class.

Typically, the SPV will have all but the most junior tranche rated by one or more of the credit rating agencies. As part of the rating process, the SPV will seek input from the rating agencies regarding how the securities need to be tranched for the most senior tranche to receive a rating of AAA (the highest possible rating). The senior tranche can receive AAA, even if there are no AAA assets in the SPV's pool, because it is the first to be paid and thus the last to suffer a loss. Its creditworthiness is enhanced because junior tranches insulate it from some level of losses from the SPV's underlying pool of assets.

The higher the credit rating, the lower the interest rate the SPV will need to offer on a particular tranche and vice versa. Thus, tranching provides investors with different risk/reward profiles. The basic idea is to convert a pool of financial assets with a single rating into various debt securities with ratings at, above, and below the pool's rating. This is considered desirable because demand for fixed income securities is divided between investors seeking the presumed safety of highly rated (AAA or AA) debt securities and those seeking the high returns offered by lower rated securities, with demand for highly rated securities the greatest. Through tranching, an SPV can take a pool of assets that falls in between these two points and create securities sought by both types of investors. In fact, the securities can be tranched easily so that the senior tranche is by far the largest tranche, aligning with the greater demand for highly rated securities.

Notwithstanding the highly rated nature of the top tranche of an SPV's debt securities, there is demand for credit protection on these securities. As noted above, the bulk of AIGFP's CDS portfolio was comprised of protection it wrote on what it refers to as the 'super senior' tranche of various types of asset-backed securities. AIG defines the 'super senior' tranche 'as the layer of credit risk senior to a risk layer that has been rated AAA by the credit rating agencies, or if the transaction is not rated, equivalent thereto' (AIG Annual Report 2007). On 31 December 2007, AIGFP had the net notional amount of protection outstanding on the super senior tranche of securities backed by the specified types of financial assets shown in Table 2.

Fall of AIG, Table 2 Notional value of credit default swaps issued by AIG, 2007

	Net notional amount (in $billions)
Corporate loans	230
Prime residential mortgages	149
Corporate debt/collateralized loan obligations	70
Multi-sector collateralized debt obligations	78
Total	527

Source: AIG Annual Report (2007)

Approximately $379 billion of AIGFP's portfolio (the corporate loans and prime residential mortgages CDSs) were written to provide various European financial institutions 'regulatory capital relief'. By purchasing CDSs from AIG, these institutions were able to reduce the amount of capital they were required by banking regulations to maintain against securities they held. The balance of AIGFP's CDS portfolio (the remaining $148 billion) was arbitrage motivated, meaning that the counterparties bought the protection as part of some type of arbitrage trading strategy.

AIGFP was able to amass such a large CDS portfolio in part because AIG contractually guaranteed all AIGFP payment obligations on the CDSs it wrote. In effect, AIGFP was leveraging the comfort provided to counterparties by AIG's stellar credit rating (AAA until 2005) and hundreds of billions in assets.

Obviously, AIGFP sold protection to make money. A former AIGFP senior executive characterized writing CDSs as 'gold' and 'free money' because AIGFP's risk models indicated that the underlying securities would never go into default (Mollenkamp et al. 2008). Thus, the CDSs would expire untriggered and AIGFP would pocket the premiums. These premiums averaged about 0.12% per year of CDS notional amount (Financial Crisis Inquiry Commission, 2011).

After the fact, the strategy was a disaster, but not necessarily irrational or reckless before the fact. Because almost all of AIGFP's CDSs were written on super senior tranches and losses are allocated sequentially starting with the equity tranche, a pool of loans backing the SPV's

securities could suffer substantial defaults before any losses would be incurred by the super senior tranche. If lower rated tranches absorb all the losses, meaning no losses have to be allocated to the super senior tranche, there will be no 'credit event' with respect to the super senior tranche and, therefore, no payment obligation under the CDS AIGFP wrote on the tranche. AIGFP's model had determined with 99.85% confidence that no credit event would ever occur with respect to the super senior tranches on which AIG wrote protection 'even in an economy as troubled as the worst post-World War II recession' (Financial Crisis Inquiry Commission 2011). This proved to be largely correct, but, as discussed next, it was not the occurrence of 'credit events' that crippled AIG, but collateral calls.

AIG's Collapse

AIG collapsed largely because of the collateral posting obligations with respect to $61.4 billion notional amount of CDSs that AIGFP wrote on debt securities with subprime mortgage exposure. As discussed above, these obligations are a common feature of CDSs designed to reduce the counterparty credit risk assumed by a CDS protection buyer. In this case, the obligations were based on (1) the difference between the notional amount of the particular CDS and the market value of the underlying debt security, and (2) the rating on the debt securities. Accordingly, as the housing market steadily declined in 2008, causing subprime borrowers to default on their mortgages, the value and ratings of the debt securities underlying the $61.4 billion of CDSs plummeted. As a result, AIG was obligated to post more and more cash collateral. By June 2008, AIG had posted $13.2 billion, and counterparties were demanding an additional $9.2 billion.

Adding to AIG's cash struggles was its securities lending program, a program managed by AIG Investments, AIG's institutional asset management unit. Under the program, AIG Investments loaned securities from the investment portfolios of AIG's insurance companies to various financial institutions (the typical reason that an institution borrows securities is to sell them short) in exchange for cash collateral posted by the borrower. AIG Investments would then invest the collateral in debt securities to earn a return which would serve as compensation for lending securities. At one point, AIG investment had loaned $76 billion in securities to US companies.

As borrowers received news about AIG's troubles, they became concerned about the safety of the cash collateral they had posted with AIG Investments. Thus, many of them decided to return lent securities and get their collateral back. Unfortunately, AIG Investments had invested a significant portion of the cash in residential mortgage-backed securities which had plummeted in value and liquidity. As a result, the program lacked sufficient funds to satisfy collateral-return obligations. Accordingly, AIG was forced to transfer billions in cash to the program, cash which was immediately paid out to these borrowers. By late August, AIG had transferred $3.3 billion in cash to the program, and borrowers were demanding billions more.

By early September 2008, AIG realized that its cash situation was dire and therefore accelerated its ongoing efforts to raise additional capital. It held discussions with private equity firms, sovereign wealth funds and other investors, but was unable to strike a deal. Furthermore, several of AIG's subsidiaries were unable to roll over their commercial paper financing, meaning that AIG was essentially shut out of the commercial paper market.

On 15 September 2008, the credit rating agencies downgraded AIG's long-term debt rating. This downgrade triggered in excess of $20 billion in additional collateral obligations because the collateral posting provisions contained in many of AIGFP's CDSs also took into account the credit rating of AIG, with a credit downgrade triggering additional posting obligations.

The day after the downgrade, AIG made a last ditch effort to raise additional financing. Among other things, AIG management met with representatives of Goldman, Sachs & Co., J. P. Morgan and the Federal Reserve Bank of New York (NY Fed) to discuss putting together a $75 billion secured lending facility syndicated among various

financial institutions. By the early afternoon, however, it was apparent that no private sector lending facility was forthcoming and that AIG 'had an immediate need for cash in excess of its available liquid resources' (AIG Quarterly Report, September 2008). AIG still had close to $1 trillion in assets but they were either illiquid or held by regulated insurance subsidiaries and thus were out of AIG's reach. As a result, the government decided to intercede.

The Bailout

On 16 September 2008, the Federal Reserve Board (Fed) announced, with the support of the US Department of the Treasury (Treasury), that it had authorized the NY Fed to rescue AIG through an $85 billion revolving credit facility (Fed Credit Facility). According to the Fed, the bailout was necessary because 'in current circumstances, a disorderly failure of AIG could add to already significant levels of financial market fragility and lead to substantially higher borrowing costs, reduced household wealth, and materially weaker economic performance' (Fed Press Release 2008). The intent of the loan was to 'facilitate a process under which AIG will sell certain of its businesses in an orderly manner, with the least possible disruption to the overall economy' (Fed Press Release 2008). In exchange for making the loan, the US government received a 79.9% equity stake in AIG.

The Fed Credit Facility kept AIG out of bankruptcy but it did not cure its financial woes. Thus, in November 2008, the government restructured its aid to AIG. The restructuring consisted of three components: an equity purchase, changes to the Fed Credit Facility, and creation of additional lending facilities. Under the equity purchase component, the US Treasury invested $40 billion in AIG under the Troubled Asset Relief Program (TARP) included in the Emergency Economic Stabilization Act of 2008. AIG used this money to pay down the Fed Credit Facility. The Fed Credit Facility was reduced from $85 billion to $60 billion and its term changed from two years to five years. To address continuing problems related to AIG's securities lending program, the NY Fed, purchased $39.3 billion face amount in residential mortgage-backed securities from AIG for $19.8 billion. These securities were purchased by AIG with cash collateral posted by borrowers under its securities lending program. AIG used the proceeds from the NY Fed and additional funds to repay this cash collateral, and it then terminated its securities lending program. Finally, to address AIG's continuing collateral posting obligations from its CDS portfolio, AIG and the NY Fed established a facility to purchase from counterparties the debt securities underlying the problematic $61.4 billion in CDSs in exchange for these counterparties concurrently terminating the related CDSs. The NY Fed provided a $30 billion term loan to fund the purchase of the CDOs, and AIG contributed $5 billion.

In March 2009, the government added an equity capital commitment facility to the aid package. Under this facility, Treasury agreed to provide AIG with up to approximately $30 billion over the ensuing five years. This last facility brought US government commitments to AIG to $182.5 billion, with AIG ultimately drawing down approximately $126.1 billion of the total.

In January 2011, the US government and AIG closed on a recapitalization plan. Under the plan, (1) AIG repaid amounts it owed under the Fed Credit Facility, (2) the various types of AIG preferred shares issued to the US government in connection with the bailout and restructuring were converted into 1.655 billion shares of AIG common stock, all of which are now held by Treasury, and (3) AIG issued Treasury approximately $20 billion of preferred equity interests in two AIG subsidiaries. Upon completion of the recapitalization, Treasury owned approximately 92% of AIG's common stock.

Regulatory Gap and Response

AIGFP was able to amass its huge portfolio of CDSs in part because of deliberate regulatory gaps. Specifically, the Commodity Futures Modernization Act of 2000 (CFMA) amended the federal securities laws to essentially prohibit the

US Securities and Exchange Commission (SEC) and the Commodity Futures Trading Commission (CFTC) from regulating over-the-counter (OTC) derivatives. Prior to CFMA passage, there was uncertainty as to whether SEC and CFTC regulations applied to OTC derivatives. A November 1999 report from a working group comprised of the Secretary of the Treasury, Chairman of the Fed, Chairman of the SEC and Chairman of the CFTC concluded that this uncertainty, 'if not addressed, could discourage innovation and growth of these important markets and damage U.S. leadership in these arenas by driving transactions off-shore' (President's Working Group on Financial Markets, 1999). Hence Congress resolved the uncertainty by making it clear that SEC and CFTC regulations did not apply. The justification for this approach was that CDSs and the like were transacted only by sophisticated parties who can fend for themselves and therefore do not need the protections afforded by SEC and CFTC regulations.

Additionally, although CDSs have characteristics of insurance contracts, they generally have not been considered insurance for purposes of state insurance regulations, and therefore have not been subject to these regulations. This was made crystal clear by the state of New York in 2004 when it amended its insurance laws specifically to exclude CDSs from coverage. A number of other states have done likewise. The basic justification for the exclusion is that the purpose of insurance regulation is to protect American consumers. Because the CDS market is comprised entirely of institutional investors, the thinking went that there was no consumer interest with respect to CDSs in need of protection.

While CDSs themselves were not regulated, many of the players in the CDS market were. For example, nationally chartered banks are supervised by the Office of the Comptroller of the Currency (OCC), and bank holding companies are regulated by the Fed. In fact, since 1999, when AIG organized AIG Federal Savings Bank, it had been subject to Office of Thrift Supervision (OTS) regulation, examination, supervision and reporting requirements. According to AIG, '[a]mong other things, this permits the OTS to restrict or prohibit activities that are determined to be a serious risk to the financial safety, soundness or stability of AIG Federal Savings Bank' (AIG Annual Report 2007). While the OTS was aware of AIG's CDS business, reviewed some of the contracts, and knew about the collateral posting provisions, they failed to recognize the extent of the risk. Congress abolished the OTS in 2010 and transferred the bulk of its responsibilities to the OCC.

In sum, because CDSs fell within a regulatory gap and the OTS did not appreciate their risks, AIGFP was able to pursue a multi-billion dollar CDS business free from regulatory filings, mandated capital requirements and government intervention.

Congress closed the CDS regulatory gap through the Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010. Among other things, the Act (1) authorizes the SEC and CFTC to regulate over-the-counter derivatives, (2) requires certain formerly OTC derivatives to be exchange-traded and centrally cleared, and (3) allows regulators to impose capital and margin requirements on swap dealers and major swap participants. As of this writing, regulations implementing these provisions are in the process of being finalized.

Conclusion

AIG collapsed because collateral obligations embedded in the CDSs it wrote triggered a chain reaction that drained it of cash. Unable to raise funds in the private markets or quickly sell off some of its trillion dollars in assets, AIG was forced to accept a government bailout. In hindsight, it is easy to conclude that AIG should have never gone into the CDS business, or at least not written the $61.4 billion of CDSs on multi-sector CDOs with subprime mortgage loan exposure. Ultimately, however, AIG took a calculated business risk that turned out disastrously.

In the wake of the bailouts of Bear Stearns, Freddie Mac and Fannie Mae, and the bankruptcy of Lehman Brothers, the government determined that the financial markets were too fragile to absorb an AIG bankruptcy. Thus, it rescued AIG with a package that soon grew to $182.5 billion.

AIGFP was able to amass its huge CDS portfolio without setting aside capital reserves or hedging its exposure because of a deliberate regulatory gap. This gap has since been closed, so a repeat of AIG is unlikely.

See Also

- ▶ Credit Crunch Chronology: April 2007–September 2009
- ▶ Fannie Mae, Freddie Mac and the crisis in US mortgage finance
- ▶ Run on Northern Rock

Bibliography

AIG Annual Report. 2007.

AIG Quarterly Report, September, 2008.

Financial Crisis Inquiry Commission. 2011. *The financial crisis inquiry report*.

Mollenkamp, C. et al. 2008. Behind AIG's fall, risk models failed to pass real-world test. *Wall Street Journal*, 3 November, at A1.

President's Working Group on Financial Markets 1999. *Report of the president's working group on financial markets, over-the-counter derivatives markets and the commodity exchange act*.

The Federal Reserve Board. 2008. Press release, 16 September.

Falling Rate of Profit

Walter Eltis

Adam Smith, David Ricardo, Karl Marx, John Stuart Mill and John Maynard Keynes all expected the rate of profit to decline in the longest of long runs. It goes without saying that their reasons differ, with the result that we have several theories which point to this possibility.

Adam Smith

Smith is generally regarded as an optimist who saw more potential for progress in real wages and labour productivity than several of his successors. He expected productivity to be constant in agriculture, while in his *Lectures* he claimed that the division of labour in industry would permit twenty million workers to produce one hundred times the output of two million (p. 392). Capital accumulation would inevitably lead to population growth which should enable these potential benefits from the division of labour to be realized, and the extra population would allow the division of labour to be further extended and permit yet higher productivity. If productivity is constant in agriculture and rising in industry, its average in industry and agriculture together, Q_y, will have a persistent tendency to rise (as Hollander (1973) shows).

At first sight most of the benefits from this rising productivity trend should go to profits and rents. The *level* of wages will be higher in a fast than in a slow-growint economy, but Smith does not expect the *high* wages of a fast growing economy to rise each year. There will be one particular *level* of wages, W, in a stationary state, a higher *level*, $W+$, in a slow growing economy, and a still higher *level,* $W++$, where capital and population are growing rapidly. As capital and employment can grow rapidly without any need for wages to rise above $W++$, all the gains in Q_y, can be added to profits and rent. This ought to produce a *rising* rate of profit, for if Q_y is rising while the wage is stuck at $W++$, then the surplus for profits and rent per worker, $(Q_y - W++)$, and the share of profits and rent in output, $(Q_y - W++)/Q_y$ will all the time increase. Unless rents take an ever growing share of 'profits and rents', this continual rise in the proportion of output which can go to profits and rent should allow the share of profits and therefore the rate of profit to keep on rising. So it is at first sight puzzling that Smith should insist in *The Wealth of Nations* (1776) that:

> In a country which had acquired that full complement of riches which the nature of its soil and climate, and its situation with respect to other countries allowed it to acquire; which could, therefore, advance no further, and which was not going backwards, both the wages of labour and the profits of stock would probably be very low (p. 111).

The cause of the declining rate of profit which takes Smith's economy gradually to a stationary

state where wages and profits are both low is most easily understood if (to follow Eltis 1984) attention is focused on agriculture, and the corn harvest in particular.

Smith suggests that each corn harvest is produced with unchanging labour productivity, 'In every different stage of improvement. . . the raising of equal quantities of corn in the same soil and climate, will, at an average, require nearly equal quantities of labour' (p. 206), while the wage is also just sufficient to buy a given quantity of corn for,

> the money price of labour . . . must always be such as to enable the labourer to purchase a quantity of corn sufficient to maintain his family either in the liberal, moderate, or scanty manner in which the advancing, stationary or declining circumstances of the society oblige the employer to maintain him (p. 509).

In a rapidly progressing economy where the wage is $W++$, this will be a sufficient sum of money to purchase a fixed quantity of corn of W_a++. If the constant output of corn per worker is Q_a, while the wage represents W_a++ of corn, the surplus that is available for profits and rent will be the constant $(Q_a - W_a++)$ of corn per worker. Therefore, if we measure output per worker and the wage in corn, there is no tendency for profits plus rent per workers to rise. If this constant share of surplus is divided equally between profits and rents, then Smith would predict an approximately constant *share* of profits in agriculture.

Smith envisages that an economy will become increasingly capital intensive.

> As the division of labour advances. . . in order to give constant employment to an equal number of workmen, an equal stock of provision, and a greater stock of materials and tools than what would have been necessary in a ruder state of things must be accumulated beforehand (p. 277).

In the case of agriculture this increase in capital intensity takes the form of a growing use of oxen ('labouring cattle') and increasing sums will be spent on fertilization and improvements to the soil. So there will be a continual tendency for the agricultural capital–output ratio to rise. With a constant share of profits (P/Y), and a rising capital-output ratio (K/Y), the rate of profit ((P/K) which is $(P/Y) \div (K/Y)$) will tend to fall. Entrepreneurs can choose whether to deploy their capital in agriculture, industry, or commerce, so the rate of profit cannot fall in agriculture without similar falls elsewhere. Hence as the agricultural rate of profit falls, the capital withdrawn from agriculture will be transferred, and the increase in competition that this causes will also force industrial and commercial profits down for,

> When the stocks of many rich merchants are turned into the same trade, their mutual competition naturally tends to lower its profit; and when there is a like increase of stock in all the different trades carried on in the same society, the same competition must produce the same effect in them all (p. 105).

The general fall in the rate of profit will gradually reduce capital accumulation, and as this diminishes, wages will fall from W_a++ to W_a+ and subsequently to W_a. At this lower wage, profits will recover a little, but the same cause, a rising K/Y in agriculture while P/Y is constant, will cause a resumption of the falling trend which will continue until the stationary state where wages and profits are both 'very low' is reached.

David Ricardo

During the Napoleonic Wars, high food prices caused British farmers to cultivate inferior land, and this led Ricardo and his great contemporary, Malthus, to attribute a major role to agricultural diminishing returns. The simplest representation of Ricardo's theory of income distribution which follows from this is also a 'corn-model' (as Sraffa (1951) and Eatwell (1975) suggest; Hollander (1979) dissents). Ricardo himself published a table in his initial statement of his new theory, *An Essay on the Influence of a Low Price of Corn on the Profits of Stock* (1815), where the wage, output and capital per agricultural worker are all expressed as quantities of corn. Because landlords receive no rent from marginal land, its entire corn output, Q_a, goes either to wages or profits. If the equilibrium or natural wage is fixed as a specific quantity of corn, W_a, then the equilibrium profits earned from the employment

of a marginal agricultural worker will be $(Q_a - W_a)$. If the capital required to employ him can be expressed as a quantity of corn, K_a, then the rate of profit at the margin will be $(Q_a - W_a)/K_a$. In the *Essay* table, corn output per worker, Q_a, falls as a growing demand for food forces the margin of cultivation onto inferior land; capital per worker, K_a, rises because extra transport costs are involved in farming inferior land (which is further from the market), while W_a, the natural wage expressed as a quantity of corn, is constant and independent of the extent to which inferior agricultural land has to be used. The continual tendency for marginal agricultural productivity to diminish, while the capital cost of employing an agricultural worker increases, persistently reduces the agricultural rate of profit, $(Q_a - W_a)/K_a$. As with Smith, if the rate of profit falls in agriculture, then competition must reduce it equally in industry and commerce.

Ricardo moved on from the 'corn-model' of the *Essay* table to a more general theory in *Principles of Political Economy and Taxation* (1817). There (as Hicks (1972) suggests) the natural wage is expressed as specific quantities of food-and-manufactures. The food items in this 'basket' of consumer goods become more expensive as agriculture is driven onto inferior land where more workers are needed to produce the food workers require, while the manufactured items included in the natural wage become cheaper as technical progress, the division of labour and a growing use of machinery reduce labour requirements. Ricardo believed that the tendency for food to require more labour will have a stronger influence on the real cost of the basket of goods that constitute the natural wage than the tendency for manufactures to require less. In consequence the aggregate labour required to produce wage goods rises all the time, so a marginal worker will spend a higher fraction of his week producing the wage goods that his constant wage requires. Then the fraction of his output that is surplus to wages and available for profits (*marginal* output never goes to rent) will have a continual tendency to fall, so there will be a declining trend in P/Y and in the rate of profit:

> The natural tendency of profits then is to fall; for, in the progress of society and wealth, the additional quantity of food required is obtained by the sacrifice of more and more labour. This tendency, this gravitation as it were of profits, is happily checked at repeated intervals by the improvements in machinery, connected with the production of necessaries, as well as by discoveries in the science of agriculture which enable us to relinquish a portion of labour before required, and therefore to lower the price of the prime necessary of the labourer (p. 120).

In this statement of Ricardo's argument in the *Principles*, the rate of profit is influenced by developments in both agriculture and industry (as Hollander (1979) emphasizes), for anything which causes workers to spend a higher fraction of time producing wage goods must increase the proportion of marginal production that goes to wages, while any increase in productivity in the manufacture of industrial *necessities* will reduce the proportion of workers' time required to produce wage goods, and so increase the fraction which can go to profits. If the tendency for real agricultural productivity to fall has more influence than the tendency for industrial productivity to rise, then P/Y, the fraction of marginal production which is surplus to wages will have a continual tendency to fall. In the *Principles* Ricardo does not repeat the proposition (from the *Essay*) that capital per worker rises as the margin of cultivation moves onto inferior land, so the tendency for the rate of profit to fall is dominated by the influence of declining agricultural productivity upon P/Y, while K/Y plays a neutral rôle.

In the *Principles* as in Smith, wages fall (from $W++$ to $W+$ and then to W) as capital accumulation and population growth diminish, and (as Hicks and Hollander (1977) show) this reduces the rate at which profits decline, without affecting the proposition that they must fall eventually to the minimum stationary state level.

In 1820 five years after the conclusion of the Napoleonic Wars, Ricardo wrote an essay for the *Encyclopedia Brittanica* on 'Funding Systems' (which Dobb (1973) considers significant) in which he modified the proposition that declining agricultural productivity *in an individual country* will inevitably cause a continual decline in its rate of profit. A country such as Britain could avoid

the influence of agricultural diminishing returns by importing its marginal food and paying with exports of manufactures:

> a country could go on for an indefinite time increasing in wealth and population, for the only obstacle to this increase would be the scarcity, and consequently high value, of food and other raw produce. Let these be supplied from abroad in exchange for manufactured goods, and it is difficult to say where the limit is at which you would cease to accumulate wealth and to derive profit from its employment (p. 179).

Ricardo did not go on to say, though it is implicit in this statement, that global diminishing returns would force profits down in the end. If marginal productivity fell at a world level, food and necessary minerals would only be obtainable at a rising real marginal cost, and wages would absorb a growing fraction of marginal production and leave a diminishing fraction over for profits, so that *P/Y* would persistently fall. A country importing food in such circumstances would face deteriorating terms of trade, and wages would have to rise in its manufacturing industries to pay the ever rising cost of imported food with the result that wages would absorb an increasing fraction of the revenues that manufacturers obtained, and so force *P/Y* downwards in precisely the manner set out in Ricardo's *Principles*.

John Stuart Mill

Mill went on to develop the economic analysis of Smith and Ricardo (as Hollander (1985) shows). He agreed that the rate of profit will be strongly influenced by population growth, capital accumulation and techniques of production which he refers to as 'the arts of production', and that these will generally advance together. But 'Agricultural skills are of slow growth', and inventions occur only occasionally, so that, as with Ricardo, agricultural improvements are no more than an intermittent counteracting tendency which temporarily relieves the adverse pressure of growing population on agricultural productivity. 'The economical progress of a society constituted of landlords, capitalists, and labourers, tends to the progressive enrichment of the landlord class; while the cost of the labourer's subsistence tends on the whole to increase, and profits to fall' (1848, pp. 731–2).

The fall in the rate of profit will continue until an eventual stationary state is reached. The minimum to which the rate of profit will then fall will be made up of two elements. There must first be a sufficient reward for the postponement of consumption to ensure the maintenance of the capital stock. This will determine the riskless rate of interest that lenders will receive from financially sound governments. The rate of profit will exceed this minimal interest rate for there will inevitably be risks of default in commercial undertakings and entrepreneurs must earn more than the rates at which they borrow if they are to be persuaded to organize production in circumstances where each faces risk. Mill believed that the minimum rate of profit set by these considerations will have a tendency to fall because a growing security of property rights would continually improve incentives to accumulate and at the same time reduce the risks involved:

> a change which has always hitherto characterized, and will assuredly continue to characterize the progress of every civilized society, is a continual increase of the security of person and property. The people of every country in Europe, the most backward as well as the most advanced, are, in each generation, better protected against the violence and rapacity of one another, both by a more efficient judicature and police for the suppression of private crime, and by the decay and destruction of those mischievous privileges which enabled certain classes of the community to prey with impunity upon the rest. They are also, in every generation, better protected, either by institutions or by manners and opinion, against arbitrary exercise of the power of government (p. 707).

For these and similar reasons, 'The risks attending the investment of savings in productive employment require, therefore, a smaller rate of profit to compensate for them than was required a century ago' (p. 737). As civilization advances, mankind becomes less the slave of the moment, and more habituated to carry their desires forward into a distant future which is 'a natural result of the increased assurance with which futurity can be looked forward to' (p. 738). All this will 'diminish

the amount of profit which people absolutely require as an inducement to save and accumulate'.

Hence the minimum rate of profit required to sustain a stationary state should fall all the time. Because there has been

> a diminution of risk and increase of providence, a profit or interest of three or four per cent is as sufficient a motive to the increase of capital in England at the present day, as thirty or forty per cent in the Burmese Empire, or in England at the time of King John.

Mill actually envisaged a time when this minimal interest rate might fall as low as one per cent.

He believed that opulent societies like 19th-century England were continually close to this minimum. If all British saving was suddenly invested at home, 'Few persons would hesitate to say, that there would be great difficulty in finding remunerative employment every year for so much new capital', and 'if the present annual amount of savings were to continue, without any of the counteracting circumstances which now keep in check the natural influence of those savings in reducing profit, the rate of profit would speedily attain the minimum, and all further accumulation of capital would for the present cease' (p. 741).

Counteracting tendencies which prevent the rate of profit from actually attaining the minimum are the diversion of a good deal of saving overseas where a higher rate of profit can be earned, and technical progress in the manufacture of wage goods which adds new opportunities for profitable investment, but Mill believed that the adverse influence on the rate of profit of the pressure to accumulate would exercise the dominant influence. Diminishing returns would even set in in North America, for as its population rose 'unless great improvements take place in agriculture' there would need to be increases in capital per worker which would gradually produce the same effects on profitability as in Europe (p. 745).

Like Ricardo, Mill did not envisage that technical progress in the production of workers' necessities would be sufficient to overcome the influence of population growth and agricultural diminishing returns, so profits would continually fall towards the level set by the returns which savers and entrepreneurs must receive, which would itself diminish.

Karl Marx

Marx did not follow Ricardo and Mill in attributing particular significance to agricultural diminishing returns. In *Capital* the production of food is not singled out in relation to the other goods that workers buy, and there is no tendency for the real cost of workers' consumer goods to rise. Marx actually argued the contrary, for he attributed great significance to the favourable effects of industrial mechanization and the division of labour. Because of these, there is a falling trend in the real cost of the goods workers buy in order to achieve the equilibrium wage, 'the value of labour in exchange'. Analogously with Smith and Ricardo, this has to provide a standard of living sufficient to sustain the population – or as Marx puts it, 'to ensure the reproduction of the working class'.

> Measured in hours of labour time, his preferred unit of value, each worker labours for $(V + S)$ hours a day, of which V suffice for the production of the wage goods required for the equilibrium wage, while the product of the remaining S hours is surplus to workers' subsistence requirements and belong to the capitalist employers. Marx describes the ratio of S, the total hours workers labour for others, to V, the hours they labour for their own subsistence needs, as 'the rate of exploitation'. Because of continuing productivity growth as a result of increasing mechanization and extensions of the division of labour, workers' subsistence needs can be met in fewer hours, so V has a persistent tendency to fall. As Marx sees no tendency for total hours of work to fall, S can rise as V falls with the result that there is a persistent tendency for S/V, the rate of exploitation, to rise. As S/V rises, so will the ratio of profits to wages and therefore the share of profits in output. Given this prediction of a rising P/Y, Marx can only arrive at the conclusion that P/K, the rate of profit, has a persistent tendency to fall, if K/Y, the capital-output ratio, rises still more persistently than P/Y.

Marx believed that there are strong historical tendencies for capital per worker and the capital-output ratio to rise. The total capital tied up in the employment of a worker consists of means of

production, namely physical capital equipment and raw materials, of C, and advance payments of wages of V per worker, in return for which the employer obtains the worker's 'labour power'. Total capital per worker is $(C + V)$, and Marx refers to C, raw materials and machinery, as constant capital, and V, the advance purchase of 'labour power', as variable capital. He believed that there is a persistent tendency for C/V which he refers to as 'the organic composition of capital' to rise. As the division of labour advances, a 'greater mass of raw material and auxiliary substances enter into the labour process', while increases in the mass of machinery, furnaces, means of transport and the means of production concentrated in buildings, are 'a condition of the increasing productivity of labour'. A 'growing extent of the means of production, as compared with the labour-power incorporated with them, is an expression of the growing productiveness of labour', and the 'law of the progressive increase in constant capital, in proportion to the variable, is confirmed at every step… whether we compare different economic epochs or different nations in the same epoch' (Vol. 1, pp. 583–4). Now the rate of profit is the ratio of total profit, that is, surplus-value, S, to total capital, $(C + V)$, and $S/(C + V)$ can be written as $(S/V)/(C/V + 1)$. The continual tendency for C/V, the organic composition of capital, to rise will all the time reduce the rate of profit, but the tendency for S/V, the rate of exploitation, to rise, will continually raise the rate of profit. Marx believed he had demonstrated a continual tendency for the rate of profit to decline, but (as Meek shows) there is no presumption that $(S/V)/(CV + 1)$ will decline if there are upward tendencies in both the organic composition of capital (C/V), and the rate of exploitation (S/V).

But Marx's conclusion of a declining rate of profit can be established if these trends are pushed to their ultimate limits. The upper limit to total surplus value per worker, S, cannot exceed one working day, while the upper limit to C, constant capital per worker (or the 'dead labour' with which workers are equipped) can become indefinitely high if the tendency for C/V to rise is continual. Thus, if the historical tendency is for S to rise to the maximum hours in a working day, S_{max}, and for C to rise without limit, then the rate of profit, $S_{max}/(C + V)$, will become indefinitely small.

Several modern commentators (e.g., Fine and Harris 1976, and Shaikh 1978) have underlined this interpretation, by adding that the upward boundary to $S/(C + V)$, which is set by the profit rate where the wage (V) is zero is S_{max}/C, which will fall continually as C rises. If the *upper limit* to the rate of profit has a continual tendency to fall, then it is a reasonable presumption that there will be a declining trend in the actual rate of profit, despite fluctuations associated with vicissitudes in wage bargaining.

Marx himself emphasized that his 'law of the tendency of the rate of profit to fall' is no more than a *tendency* which can and will be counteracted by a variety of developments over considerable periods. Wage costs may fall for a time and permit the rate of profit to rise if imported workers' consumer goods can be produced more cheaply overseas. New industries may begin to produce with low capital intensity (and therefore a low C/V): in Marx's words they begin by employing mainly living labour. But as these industries develop, capital intensity will rise and the ratio of dead to living labour increase, so that C/V rises in the same way as in older industries. Another possibility is that capital equipment may fall in price relative to consumer goods, and in this case industry will become more capital intensive in technical terms without any necessary tendency for the organic composition of capital to rise. The 'technical composition of capital' (C/V measured in technical units) would still be rising, but not its 'organic composition' which is C/V in Marx's labour units.

But the tendency for growing capital intensity to reduce the rate of profit would dominate any secular trend, for these helpful developments could only operate for a time.

John Maynard Keynes

There is an echo of Mill's theory in *The General Theory of Employment, Interest and Money* (1936). Keynes believed that if a country could

reduce its rate of interest to a level compatible with full employment, and then invest its full employment saving, it would be 'comparatively easy to make capital-goods so abundant that the marginal efficiency of capital is zero'. He believed that 'a properly run community equipped with modern technical resources, of which the population is not increasing rapidly, ought to be able to bring down the marginal efficiency of capital in equilibrium approximately to zero within a single generation' (pp. 220–21).

Thus Keynes, like Mill, believed that a modern economy's potential to accumulate greatly transcended the rate at which new investment opportunities would arise, with the result that the rate of profit would rapidly fall towards the stationary state level if its full potential for accumulation could ever be realized.

Conclusion

The theories of these great economists have rested on three general predictions about the future development of capitalist economies which have not been borne out empirically.

Smith and Marx both believed that capital would have a persistent tendency to grow faster than output, which could be expected to produce a declining tendency in the rate of profit, and the trend in the capital-output ratio was indeed upward prior to 1776 and 1867. But the British capital-output ratio has been approximately stable since 1867 (Matthews et al. 1982), while the United States capital-output ratio has been falling (Klein and Kosobud 1961). Few now speak of a long term tendency for the capital-output ratio to rise, so this line of argument finds little echo in 20th-century economics.

Ricardo and Mill were much influenced by a belief that the adverse influence of agricultural diminishing returns would inevitably outweigh any favourable effects from technical progress, with the result that the real cost of workers' necessities would rise continuously and squeeze the rate of profit. But since they wrote, there has been no tendency for the real cost of food and raw materials to rise faster than manufactures. The terms of trade have fluctuated a good deal, but technical progress has raised productivity enormously in both industry and agriculture, and there has been no tendency for a rising relative cost of food to squeeze profits in the manner that Ricardo and Mill expected. Some futurologists predict a gradual depletion of the world's natural resources with inevitable Ricardian (and Malthusian) consequences, but the 20th century itself has provided no empirical support for their pessimism.

Mill and Keynes were impressed by the proposition that continuing capital accumulation would exhaust opportunities for profit faster than new investment opportunities can be created. But since World War II technical progress has accelerated, and there have been decades when new investment opportunities providing enormous scope for profitable investment have emerged. It is rarely argued now that there is any *necessary* tendency for the new investment opportunities created by technical advance to fall short of actual investment so that the marginal efficiency of investment must tend to fall.

So there is little late 20th-century support for the theories which have been outlined. There is however a further hypothesis which is germane to the general direction of Marx's political and social thought. If there is a continual increase in the power of workers in wage bargaining in comparison with the power of capitalists to resist their influence, then the share and rate of profit will have a tendency to fall. This will be reinforced if workers' political representatives exercise a growing legislative influence over wage bargaining and price formation. If workers become immune from dismissal or redundancy without compensation, while the prices companies set are increasingly subject to public scrutiny, then there will be an accompanying tendency for the rate and share of profits to decline. In the 1970s in several countries the political power of the working class appeared to rise with accompanying shifts in income distribution, but political developments have been in the other direction in the early 1980s, so as with previous hypotheses, there is no particular reason to anticipate any clear future trend.

See Also

- ▶ British Classical Economics
- ▶ Marxist Economics
- ▶ Stationary State

Bibliography

Primary

Keynes, J.M. 1936. *The general theory of employment, interest and money*. Republished as vol. VII of *The collected writings of John Maynard Keynes*, London: Macmillan, 1973.

Marx, K. 1867–94. *Capital*, 3 vols. Reprinted, Moscow: Progress Publishers for Lawrence and Wishart, 1974.

Mill, J.S. 1848. *Principles of political economy with some of their applications to social philosophy*, 2 vols. Reprinted in *Collected works of John Stuart Mill*, vols II and III, ed. J.M. Robson, Toronto: University of Toronto Press, 1965.

Ricardo, D. 1815. *An essay on the influence of a low price of corn on the profits of stock*. Reprinted in vol. IV of *The works and correspondence of David Ricardo*, 11 vols, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951–73.

Ricardo, D. 1817. *On the principles of political economy and taxation*. Reprinted as vol. I of *The works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951– 73.

Ricardo, D. 1820. Funding systems. In *Encyclopedia Brittanica*, Supplement to 4th ed. Reprinted in vol. IV of *The works and correspondence of David Ricardo*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Smith, A. 1763. In *Lectures on jurisprudence*, ed. R.L. Meek, D.D. Raphael, and P.G. Stein. Oxford: Oxford University Press, 1978.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, 2 vols. Reprinted, ed. R.H. Campbell, A.S. Skinner and W.B. Todd. Oxford: Oxford University Press, 1976.

Secondary

Dobb, M. 1973. *Theories of value and distribution since Adam Smith*. Cambridge: Cambridge University Press.

Eatwell, J. 1975. The interpretation of Ricardo's *Essay on profits*. *Economica* 42(166): 182–187.

Eltis, W. 1984. *The classical theory of economic growth*. London: Macmillan.

Fine, B., and L. Harris. 1976. Controversial issues in Marxist economic theory. *Socialist Register*.

Hicks, J. 1972. *Ricardo's theory of distribution*. Reprinted in *Classics and moderns*. Oxford: Blackwell, 1983.

Hicks, J., and S. Hollander. 1977. Mr Ricardo and the moderns. *Quarterly Journal of Economics* 91: 351–369.

Hollander, S. 1973. *The economics of Adam Smith*. Toronto: University of Toronto Press.

Hollander, S. 1979. *The economics of David Ricardo*. Toronto: University of Toronto Press.

Hollander, S. 1985. *The economics of John Stuart Mill*, 2 vols. Oxford: Blackwell.

Klein, L.R., and R.F. Kosobud. 1961. Some econometrics of growth: Great ratios of economics. *Quarterly Journal of Economics* 75(2): 173–198.

Matthews, R.C.O., C.H. Feinstein, and J.C. Odling-Smee. 1982. *British economic growth, 1856–73*. Oxford: Oxford University Press.

Meek, R.L. 1960. The falling rate of profit. *Science and Society* 24: 36–52.

Shaikh, A. 1978. Political economy and capitalism: Notes on Dobb's theory of crisis. *Cambridge Journal of Economics* 2: 233–251.

Sraffa, P. 1951. Introduction to D. Ricardo. In *On the principles of political economy and taxation*. Cambridge: Cambridge University Press.

Falsificationism

Daniel M. Hausman

Keywords

Blaug, M; Falsificationism; Friedman, M; Popper, K; Scientific method; Theory appraisal; Verification

JEL Classification

D6

Many economists would emphasize that scientific claims must be capable of *falsification*. According to Milton Friedman, an hypothesis 'is rejected if its predictions are contradicted.... Factual evidence can never "prove" a hypothesis; it can only fail to disprove it...' (1953, p. 9). These claims echo Karl Popper's philosophy of science, which, on one interpretation, maintains that what distinguishes scientific theories from theories that are not scientific is that scientific theories are falsifiable. A theory is falsifiable if it is logically inconsistent with some finite set of 'basic statements' – that is, true or false reports of

observation. A true theory will not be inconsistent with any set of true basic statements, but it will still be falsifiable because it is inconsistent with (or 'forbids') some observation reports. In other words, logic and observation can force one to give up falsifiable theories. Popper notes that there is an asymmetry between falsification and verification: basic statements can be logically inconsistent with universal generalizations and can thereby disprove them, but they do not imply that any universal generalizations are true. In his view scientific knowledge grows exclusively from falsification. Verification and even confirmation are impossible.

Although Popper distinguishes theories that are falsifiable from theories that are not falsifiable, he is also distinguishes the 'critical' attitudes and norms that characterize scientists – who are willing to test theories harshly and to give up claims that do not pass the test – from the dogmatic attitudes of non-scientists, who seek supporting evidence and explain away apparently disconfirming evidence. It is this latter methodological distinction between science and non-science that is Popper's more important contribution.

To maintain that scientific theories are falsifiable is problematic, because, with very few exceptions, scientific theories are not testable or falsifiable by themselves. Observing an increase in demand for some commodity after a rise in its price does not falsify the law of demand if there has been a change in tastes, an even greater increase in the price of a close substitute, a general rise in the price level and hence a drop in the real price, or some other complicating factor. To say that an hypothesis 'is rejected if its predictions are contradicted' is misleading, because hypotheses rarely have predictions of their own. Significant scientific hypotheses imply predictions only when combined with other statements. So, if one insists that scientific claims have to be testable all by themselves, virtually nothing in science counts as science. On the other hand, if one insists only that, like the law of demand, scientific claims must be falsifiable in combination with other claims, then one cannot rule out even the most blatant pseudo-sciences. When Popper criticizes the scientific credentials of Freudian psychology, he does not maintain that, coupled with other statements, it makes no predictions. His criticism is instead that, when those predictions fail, psychoanalysts never cast blame on Freud's theory.

What distinguishes sciences from pseudo-sciences is methodology: when amalgams of theories and various auxiliary hypotheses make false predictions, scientists, unlike practitioners of pseudo-science, are willing to modify or even discard their theories. However, it is difficult to specify exactly how willing scientists should be to surrender their theories. Deciding whether observations give one good reason to reject an hypothesis, like deciding whether observations give one good reason to accept an hypothesis, requires weighing alternative explanations of the data. There is no simple asymmetry between falsification and confirmation.

The significance of falsification is methodological rather than logical or linguistic – a question of the norms that should govern science. The message of falsification is that science treats its findings as subject to criticism and revision. How can one make this platitude concrete? As even Popper and his followers have recognized, some dogmatism may be a good thing. Theories are hard to come by and should not be surrendered too easily. What characterizes successful sciences is on the one hand a mixture of attitudes on the part of individual scientists, with some much more critical than others, and on the other hand an institutional structure in which criticism is not too risky to individuals, and successful criticisms are strongly rewarded.

Those commentators on economic methodology who have been most influenced by Popper have generally been critical of economists. Mark Blaug, for example, argues that economists practise 'innocuous falsificationism' (1976, pp. 159–60), paying lip service to the importance of falsification while in fact showing little interest in criticism.

See Also

► Theory Appraisal

Bibliography

Blaug, M. 1976. Kuhn versus Lakatos *or* Paradigms versus research programmes in the history of economics. In *Method and appraisal in economics*, ed. S. Latsis. Cambridge, UK: Cambridge University Press.
Caldwell, B. 1991. Clarifying popper. *Journal of Economic Literature* 29: 1–33.
Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*. Chicago: University of Chicago Press.
Popper, K. 1957. *The poverty of historicism*. New York: Harper and Row.
Popper, K. 1966. *The open society and its enemies*, vol. 2, 5th ed. Princeton: Princeton University Press.
Popper, K. 1968. *The logic of scientific discovery,* Revised edn. London: Hutchinson & Co.
Popper, K. 1969. *Conjectures and refutations; The growth of scientific knowledge*, 3rd ed. London: Routledge and Kegan Paul.
Popper, K. 1972. *Objective knowledge: An evolutionary approach*. Oxford: Clarendon.
Popper, K. 1976. *The unended quest*. La Salle: Open Court.

Fama, Eugene F. (1939–)

G. William Schwert

Abstract

Eugene Fama is known as the father of empirical finance. Over an unusually active career that spans more than five decades, Fama has produced pioneering research on efficient capital markets and asset pricing models, as well as the behaviour of interest rates, exchange rates, futures prices and inflation rates. He has also produced important papers on capital structure and payout policy. His theoretical work on agency problems and banking is groundbreaking and influential. In addition, Fama's influence on finance through the doctoral students he has supervised and his diligent work as a professional colleague are widely recognised and appreciated.

Keywords

Agency costs; American finance association; Anomaly; Capital structure; Capital asset pricing model; Dimensional fund advisors; Dividend policy; Dividend yields; Efficient capital markets; Exchange rates; Fama–French three factor model; Financial markets; Governance; Inflation rates; Interest rates; Small firm effect; Stock returns; University of Chicago; Value effect

JEL Classifications

G11; G12; G14; G15; G21; G31; G32; G34; G35; E31; E43; E44; C31; C46; B31

Eugene Fama began his doctoral studies at the University of Chicago in the early 1960s when finance was first becoming the subject of scientific inquiry. The existence of computing technology and the creation of new financial databases at that time allowed Fama, his co-authors and his students to make a big leap forward in the types of questions that could be studied and the kinds of evidence that could be produced. The synergy between the new possibilities of studying financial data and the ideas that were being produced by the pioneers of financial economics at Chicago and MIT at that time led to an explosion of theories and evidence that remain the foundation for what financial economists know and study to this day. Eugene Fama led the vanguard that made finance one of the most productive and influential fields of economics.

Born on 1 February 1939 in Boston, Massachusetts, Fama graduated from Tufts University in 1960 with numerous academic and athletic awards, including honours in Romance Languages. He then entered the doctoral programme of the Graduate School of Business of the University of Chicago, receiving his MBA in 1963 and his PhD in 1964. His doctoral dissertation, 'The Behavior of Stock Prices', supervised by Merton Miller and Harry Roberts, was published in the *Journal of Business* in 1965 and is frequently cited 50 years later.

Fama joined the faculty of the GSB at Chicago and began a career of teaching and research that has spanned more than 50 years at the date of this article. He was appointed as a chaired professor in 1973 and is now the Robert R. McCormick Distinguished Service Professor of Finance.

During his career he has been honoured in many ways. He is the recipient of the 2013 Nobel Prize in Economic Sciences, along with Lars Hansen and Robert Shiller. He received the Belgian National Science Prize (1982) and honorary doctor of laws degrees from the University of Rochester (1987), DePaul University (1989), Catholic University of Leuven (1995) and Tufts University (2002). He has been elected as a Fellow of the American Finance Association (the first elected, in 2001), the Econometric Society and the American Academy of Arts and Sciences. He was the first recipient of the Deutsche Bank Prize in Financial Economics (2005), the first recipient of the Morgan Stanley American Finance Association Award for Excellence in Finance (2007) and the first recipient of the Onassis Prize in finance (2009). Many of his papers have won awards for being among the best in publications such as the *Journal of Finance* and the *Journal of Financial Economics*. He is one of the most cited authors across all fields of economics.

Efficient Capital Markets

Fama essentially invented the concept of efficient capital markets in his early work on the time series behaviour of stock prices. He extended it, in collaboration with Larry Fisher et al. (1969), in a study of stock splits that pioneered the technique of 'event studies'. They found that once information about the existence of a stock split becomes known to the public, there are no abnormal returns available by either buying or selling a stock that is splitting. Event studies have been used in many fields of applied economics and have become an integral part of securities law through the concept of 'reliance' and 'fraud on the market'.

Three subsequent papers, Fama (1970, 1991, 1998), and Chap. 5 of his 1976 book, *Foundations of Finance*, articulate the important idea that all tests of market efficiency are dependent on some assumption about 'equilibrium expected returns'. In other words, to test whether a security or trading strategy earns 'abnormal returns', it is first necessary to specify a model for 'normal returns'. Thus, while the earliest tests of market efficiency were based on things like serial correlations of stock returns, modern tests of market efficiency are based on much more sophisticated benchmarks that allow for cross-sectional and time series variation in asset returns that are assumed to represent differences in risks, liquidity or some other economic factor that would explain these differences. Thus, when tests find that some trading strategy cannot be explained by the maintained asset pricing model, it is referred to as an 'anomaly', that is something awaiting further explanation. An anomaly may represent true abnormal returns – essentially a money-making opportunity – or it may merely represent an incomplete model of the risk of that particular asset or class of assets. Anomalies represent opportunities for further exploration, not a definitive proof of market inefficiency.

Asset Pricing

At the same time that Fama was formulating the idea of the efficient markets hypothesis, Sharpe (1964), Lintner (1965) and Mossin (1966) were developing the capital asset pricing model (CAPM) based on Markowitz's (1952, 1959) model for portfolio selection. Fama (1968) clarified this model and showed that the apparent differences between the Sharpe and Lintner models were not real.

Fama and MacBeth (1973) and Black et al. (1972) performed early tests of the CAPM. These papers developed the empirical tools that have been used since that time to test more sophisticated models of asset pricing that allow for multiple sources of risk. For example, the technique of estimating cross-sectional regressions of portfolio returns on estimates of portfolio risk in each month and then using the monthly time series of these estimates to estimate the average risk premium and the standard error of the estimate has been widely used and is commonly called the 'Fama–MacBeth' technique.

Fama next turned to studying the relation between nominal interest rates and the inflation rates of consumption goods' prices. His 1975 paper in the *American Economic Review* used the simple predictive regression of inflation rates on the interest rate for that month to study the joint

hypothesis of efficient markets for Treasury bills and an expected real return to bills that is constant over time. For the 1953–71 sample period he studied, this simple model works well. An implication of this simple model is that realised real returns to Treasury bills are serially uncorrelated, even though serial correlations of nominal interest rates and inflation rates are substantially non-zero.

Fama and Schwert (1977) took the results of Fama (1975) and studied the relation between various classes of assets, including stocks, bonds, Treasury bills, real estate and human capital with the expected and unexpected components of inflation. A surprising finding that was an early part of the literature on time-varying expected stock returns was that expected stock returns were negatively related to nominal interest rates. This also meant that the excess returns of stocks relative to Treasury bills were even more negatively related to nominal interest rates. Fama and French (1988) extended the idea that expected returns to stocks vary over time using aggregate dividend yields as a predictor variable. The literature on time-varying expected returns to assets has since exploded after these early contributions.

Fama and French (1992, 1993) began a new approach to the empirical modelling of expected stock returns using firm size and book-to-market or 'value' factors in addition to the return to a market portfolio of stocks. The 'Fama–French three factor model' became the benchmark that others in both academia and Wall Street used to measure expected stock returns. The size factor builds on earlier work by Rolf Banz (1981) in his dissertation, which was supervised by Fama. In subsequent work Davis et al. (2000) showed that the three factor model works well in US data before 1962, and Fama and French (1998) showed that it works well in equity markets outside the USA.

While the academic impact of the Fama–French model is substantial (as reflected in thousands of citations to their papers), it is perhaps even more impressive that their work has had a large impact on professional practice. For example, the firm Dimensional Fund Advisors (DFA), for which Fama has been a Board member since its founding and at times was its Director of Research, has grown to have more than $380 billion under management, largely following strategies motivated by the Fama–French model. David Booth, one of the co-founders of DFA and a student of Fama, gave a naming gift to the Chicago Business School in honour of the contributions that Fama made to the success of DFA.

Recently, Fama and French (2015) have extended their research to include two additional factors that reflect evidence produced by others that the three-factor model can be improved. The new factors reflect the profitability of the firm and the rate of investment. They find that, in general, smaller firms earn higher average returns, value firms (high B/M) earn higher average returns than growth firms (low B/M), firms that are more profitable earn higher average returns and firms that invest less earn higher average returns.

Interest Rates, Exchange Rates and Futures Prices

Fama developed a method to analyse the term structure of interest rates and exchange rates that is based on the following decomposition:

$$\text{Forward Rate}_t - \text{Spot Rate}_t = \text{Premium}_t + [\text{E}(\text{Spot Rate}_{t+1}) - \text{Spot Rate}_t].$$

If Premium$t$ is constant over time, the current spread between the spot rate and the forward rate is just a forecast of the future spot rate. Based on extensive empirical analysis, he concludes that most of the variation in forward rates relative to spot rates is due to variation in premiums, so that forward rates alone are poor forecasts of future spot rates. He also finds that premiums and expected changes in spot rates are negatively correlated, although the reason for this negative correlation remains a puzzle.

Fama (1984a) and Fama and Bliss (1987) apply this analysis to the term structure of interest rates. Fama (1984b) studies forward exchange rates and Fama and French (1987) study the structure of futures prices using this approach. Even today, this approach to studying the structure of future or forward interest rates or exchange rates remains standard in the literature.

Another recent innovation in the exchange rate literature is to use factors, similar to Fama and French (1989, 1993), to help explain average currency returns (e.g. Lustig et al. (2011)).

Agency Theory

Stimulated by the Jensen and Meckling (1976) paper on agency problems, Fama (1980) explores the role that competition from internal and external managerial labour markets can play to mitigate or control agency problems within firms. He then collaborated with Mike Jensen on papers (1983a, b) that extend the Jensen–Meckling analysis to a variety of settings, including not-for-profit organisations, professional partnerships and others. All of these papers have been cited thousands of times and thus influenced many subsequent papers.

Corporate Finance and Banking

At various times during his career Fama has delved into a variety of standard topics in the corporate finance literature, including cash management models and studies of capital structure and of dividend policy. He also wrote fundamental papers on the differences between commercial banks and other kinds of financial institutions, and the implications of that for monetary policy. The Fama and Miller (1972) book is a concise and complete exposition of the Modigliani–Miller irrelevance propositions about capital structure and dividend policy. For many people, this set of papers would represent a very successful career, but for Fama these papers were an interesting subplot in his research portfolio.

Fama's Students

Fama's earliest PhD students at Chicago were a group that became the pioneers of finance and accounting. Michael Jensen, Myron Scholes, Richard Roll, Ross Watts, William Beaver and Ray Ball were all supervised by Fama and have subsequently produced research that has been cited tens of thousands of times by other authors. Later generations of students included Campbell Harvey, Brad Barber, Francis Longstaff, Robert Stambaugh and many others (including the author of this article). In total, Fama served on dissertation committees of more than 100 doctoral students at the University of Chicago Business School and Economics Department. Those students have written papers that have been cited more than 585,000 times on Google Scholar (Schwert and Stulz (2014) provide detailed information).

The Legacy

Eugene Fama, along with Merton Miller, built a very strong finance group at the University of Chicago through their intellectual leadership. Fama's devotion to intellectual honesty and the importance of careful data analysis, along with his commitment to providing comments and guidance to colleagues, set an important tone for the entire group. Similarly, his approach to research and writing are much appreciated by colleagues across the finance profession. His energy and enthusiasm for his research remains strong more than 50 years after he began his career. Fama (2011, 2014) provide more detailed and personal insights into his research career.

See Also

- ▶ Banking Industry
- ▶ Capital Asset Pricing Model
- ▶ Corporate Governance
- ▶ Efficient Markets Hypothesis
- ▶ Financial Market Anomalies
- ▶ Miller, Merton (1923–2000)
- ▶ Scholes, Myron (born 1941)
- ▶ Stock Price Predictability
- ▶ Term Structure of Interest Rates

Selected Works

Davis, J.L., E.F. Fama, and K.R. French. 2000. Characteristics, covariances, and average returns: 1929 to 1997. *Journal of Finance* 55: 389–406.

Fama, E.F. 1965. The behavior of stock market prices. *Journal of Business* 38: 34–105.
Fama, E.F. 1968. Risk, return and equilibrium: Some clarifying comments. *Journal of Finance* 23: 29–40.
Fama, E.F. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25: 383–417.
Fama, E.F. 1975. Short-term interest rates as predictors of inflation. *American Economic Review* 65: 269–282.
Fama, E.F. 1976. *Foundations of Finance*. New York: Basic Books.
Fama, E.F. 1980. Agency problems and the theory of the firm. *Journal of Political Economics* 88: 288–307.
Fama, E.F. 1984a. The information in the term structure. *Journal of Financial Economics* 13: 509–528.
Fama, E.F. 1984b. Forward and spot exchange rates. *Journal of Monetary Economics* 14: 319–338.
Fama, E.F. 1991. Efficient markets II. *Journal of Finance* 46: 1575–1617.
Fama, E.F. 1998. Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics* 49: 283–306.
Fama, E.F. 2011. My life in finance. *Annual Review of Financial Economics* 3: 1–15.
Fama, E.F. 2014. Two pillars of asset pricing. *American Economic Review* 104: 1467–1485.
Fama, E.F., and R.R. Bliss. 1987. The information in long-maturity forward rates. *American Economic Review* 77: 680–692.
Fama, E.F., and K.R. French. 1987. Commodity futures prices: Evidence on forecast power and premiums. *Journal of Business* 60: 55–73.
Fama, E.F., and K.R. French. 1988. Dividend yields and expected stock returns. *Journal of Financial Economics* 22: 3–25.
Fama, E.F., and K.R. French. 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics* 25: 23–49.
Fama, E.F., and K.R. French. 1992. The cross-section of expected stock returns. *Journal of Finance* 47: 427–465.
Fama, E.F., and K.R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33: 3–56.
Fama, E.F., and K.R. French. 1998. Value versus growth: the international evidence. *Journal of Finance* 53: 1975–1999.
Fama, E.F., and K.R. French. 2015. A five-factor asset pricing model. *Journal of Financial Economics*, forthcoming.
Fama, E.F., and M.C. Jensen. 1983a. Separation of ownership and control. *Journal of Law and Economics* 26: 301–325.
Fama, E.F., and M.C. Jensen. 1983b. Agency problems and residual claims. *Journal of Law and Economics* 26: 327–349.
Fama, E.F., and J.D. MacBeth. 1973. Risk, return, and equilibrium: Empirical tests. *Journal of Political Economics* 81: 607–636.
Fama, E.F., and M.H. Miller. 1972. *The theory of finance*. New York: Holt, Rinehart & Winston.
Fama, E.F., and G.W. Schwert. 1977. Asset returns and inflation. *Journal of Financial Economics* 5: 115–146.
Fama, E.F., L. Fisher, M.C. Jensen, and R. Roll. 1969. The adjustment of stock prices to new information. *International Economic Review* 10: 1–21.

Acknowledgments I am grateful for comments from Eugene F. Fama and René M. Stulz.

Bibliography

Banz, R.W. 1981. The relationship between return and market value of common stocks. *Journal of Financial Economics* 9: 3–18.
Black, F., M.C. Jensen, and M. Scholes. 1972. The capital asset pricing model: Some empirical tests. In *Studies in the theory of capital markets*, ed. M.C. Jensen, 79–121. New York: Praeger.
Jensen, M.C., and W.H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.
Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics Statistics* 47: 13–37.
Lustig, H., N. Roussanov, and A. Verdelhan. 2011. Common risk factors in currency markets. *Review of Financial Studies* 24: 3731–3777.

Markowitz, H.M. 1952. Portfolio selection. *Journal of Finance* 7: 77–99.
Markowitz, H.M. 1959. *Portfolio selection: Efficient diversification of investments*. New York: Wiley.
Mossin, J. 1966. Equilibrium in a capital asset market. *Econometrica* 34: 768–783.
Schwert, G. W., and Stulz, R. M. 2014. Gene Fama's impact: a quantitative analysis. In: The Fama Portfolio (eds. J. Cochrane and T. Moskowitz). University of Chicago Press, Chicago, forthcoming.
Sharpe, William F. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.

Family

Gary S. Becker

In virtually every known society – including ancient, primitive, developing, and developed societies – families have been a major force in the production and distribution of goods and services. They have been especially important in the production, care, and development of children, in the production of food, in protecting against illness and other hazards, and in guaranteeing the reputation of members. Moreover, parents have frequently displayed a degree of self-sacrifice for children and each other that is testimony to the heroic nature of men and women.

Of course, families have radically changed over time. The detailed kinship relations in primitive societies traced by anthropologists contrast with the predominance of nuclear families in modern societies, where cousins often hardly know each other, let alone interact in production and distribution. The obligations in many societies to care for and maintain elderly parents is largely absent in modern societies, where the elderly either live alone or in nursing homes.

Nevertheless, families are still much less prominent in economic analysis than in reality. Although the major economists have claimed that families are a foundation of economic life, neither Marshall's *Principles of Economics*, Mill's *Principles of Political Economy*, Smith's *Wealth of Nations* nor any of the other great works in economics have made more than casual remarks about the operation of families.

One significant exception is Malthus's model of population growth. Malthus was concerned with the relation between fertility, family earnings, and age at marriage, and he argued that couples usually do (or should) marry later when economic circumstances are less favourable. However, this important insight (see Wrigley and Schofield 1981, for evidence that prior to the 19th century, marriage rates in England did increase when earnings rose) had no cumulative effect on the treatment of the family by economists.

During the last 40 years, economists have finally begun to analyse family behaviour in a systematic way. No aspect of family life now escapes interpretation with the calculus of rational choice. This includes such esoteric subjects as why some contraceptive techniques are preferred to others, and why polygamy declined, as well as more 'traditional' subjects such as what determines age at marriage, number of children, the amount invested in the human capital of children, and the amount spent by children on the care of elderly parents. This essay sets out the 'economic approach' to various aspects of family behaviour. Detailed discussions of particular aspects can be found in the bibliography.

Fertility

Let us start with the Malthusian problem: how is the number of children, or fertility, of a typical family determined? Crucial to any discussion is the recognition, taken for granted by Malthus, that men and women strongly prefer their own children to children produced by others. This preference to produce one's children eventually helped stimulate economists to recognize that families, and households more generally, are important producers as well as consumers.

The desire for own children means that the number of children in a family is affected by supply conditions. Supply is determined by knowledge of birth control techniques, and by the capacity to produce children, as related to age, nutrition, health, and other variables.

The demand side emerges through maximization of the utility of a family that depends on the quantity of children (n) and other commodities (z), as in

$$U = U(n, z). \tag{1}$$

Utility is maximized subject not only to household production functions for children and other commodities, but also to constraints on family resources. Money income is limited by wage rates and the time spent working, and the time available for household production is limited by the total time available. These constraints are shown by the following equations where λ is the marginal utility of family income. The total net cost of rearing a child (Π_n) equals the value of the goods and services that he consumes, plus the value of the time spent on him by family members ($\Sigma w_i(t_{ni})$), minus his earnings that contribute to family resources.

$$\left.\begin{array}{l} p_n + p_z z = \sum w_i t_{wi} + \upsilon \\ t_{ni} \ + t_{zi} \ + t_{wi} \ = t \end{array}\right\} \text{all } i \in f, \tag{2}$$

where t_{wi} is the hours worked by the ith family member, w_i is his or her hourly wage, v is non-wage family income, t_{ni} and t_{zi} are the time allocated to children and other commodities by the ith member, and t is the total time available per year or other time unit.

By substituting the time constraints into the income constraint, one derives the family's full income (S):

$$\begin{aligned} (p_n + \textstyle\sum w_i t_{ni})n + (p_z + \sum w_i t_{zi})Z &= \sum w_i t + \upsilon \\ &= S, \Pi_n n + \Pi_z Z = S. \end{aligned} \tag{3}$$

If utility is maximized subject to full income, the usual first order conditions follow:

$$\frac{\partial U}{\partial n} = \lambda \Pi_n, \tag{4}$$

and

$$\frac{\partial U}{\partial z} = \lambda \Pi_z, \tag{5}$$

The basic theorem of demand states that an increase in the relative price of a good reduces the demand for that good when real income is held constant. If the qualification about income is ignored, then, in particular, an increase in the relative price of children would reduce the children desired by a family. The net cost of children is reduced when opportunities for child labour are readily available, as in traditional agriculture. This implies that children are more valuable in traditional agriculture than in either cities or modern agriculture, and explains why fertility has been higher in traditional agriculture (see the evidence in Jaffe 1940; Gardner 1973).

Production and rearing of children have usually involved a sizeable commitment of the time of mothers, and sometimes also that of close female relatives, because children tend to be more time intensive than other commodities, especially in mother's time (i.e. in equation (3), $p_n/\Pi_n < p_z/\Pi_z$). Consequently, a rise in the value of mother's time would reduce the demand for children by raising the relative cost of children. In many empirical studies for primitive, developing, and developed societies, the number of children has been found to be negatively related to various measures of the value of mother's time (see e.g. Mincer 1962; Locay 1987).

Women with children have an incentive to engage in activities that are complementary to child care, including work in a family business based at home, and sewing or weaving at home for pay. Similarly, women who are involved in complementary activities are encouraged to have children because children do not make such large demands on their time. This explains why women on dairy farms have more children than women on grain farms: dairy farming inhibits off-farm work because that is not complementary with children.

During the past one hundred years, fertility declined by a remarkable amount in all Western countries; as one example, married women in the US now average a little over two live births compared with about five-and-a-half live births in 1880 (see US Bureau of the Census 1977). Economic development raised the relative cost of children because the value of parents' time

F

increased, agriculture declined, and child labour became less useful in modern farming. Moreover, parents substituted away from number of children toward expenditures on each child as human capital became more important not only in agriculture, but everywhere in the technologically advanced economies of the 20th century (for a further discussion, see Becker 1981, ch. 5).

'Quality' of Children

The economic approach contributes in an important way to understanding fertility by its emphasis on the 'quality' of children. Quality refers to characteristics of children that enter the utility functions of parents, and has been measured empirically by the education, health, earnings, or wealth of children. Although luck, genetic inheritance, government expenditures, and other events outside the control of a family help determine child quality, it also depends on decisions by parents and other relatives.

The quality and quantity of children interact not because they are especially close substitutes in the utility function of parents, but because the true (or shadow) price of quantity is partly determined by quality, and vice versa. To show this, write the utility function in equation (1) as

$$U = U(n, q, Z), \tag{6}$$

where q is the quality of children. Also write the family budget equation in equation (3) as

$$_{n}n + \Pi_q q + \Pi_c nq + \Pi_z Z = S, \tag{7}$$

where Π_n is the fixed cost of each child, Π_q is the fixed cost of a unit of quality, and Π_c is the variable cost of children.

By maximizing utility subject to the family income constraint, one derives the following first order conditions:

$$\frac{\partial U}{\partial n} = \lambda(\Pi_n + \Pi_c q) = \lambda \Pi_n^*, \tag{8}$$

$$\frac{\partial U}{\partial q} = \lambda\left(\Pi_q + \Pi_c q\right) = \lambda \Pi_q^*, \tag{9}$$

$$\frac{\partial U}{\partial z} = \lambda \Pi \tag{10}$$

Quantity and quality interact because the shadow price of quantity (Π_n^*) is positively related to the quality of children, and the shadow price of quality (Π_q^*) is positively related to the quantity of children.

To illustrate the nature of this interaction, consider a rise in the fixed cost of quantity ($\prod_n$) that raises the shadow price of quantity (Π_n^*), and thereby reduces the demand for quantity. A reduction in quantity however, lowers the shadow price of quality (Π_q^*), which induces an increase in quality. But the increase in quality, in turn, raises further the shadow price of quantity, which reduces further the quantity of children, which induces a further increase in quality, and so on until a new equilibrium is reached. Therefore, a modest increase in the fixed cost of quantity could greatly reduce the quantity of children, and greatly increase their quality, *even when quantity and quality are not good substitutes in the utility function*.

The interaction between quantity and quality can explain why large declines in fertility are usually associated with large increases in the education, health, and other measures of the quality of children (see the evidence in Becker 1981, ch. 5). It also explains why quantity and quality are often negatively related among families: evidence for many countries indicates that years of schooling and the health of children tend to be negatively related to the number of their siblings (see e.g. De Tray 1973; Blake 1981).

The influence of parents on the quality of their children links family background to the achievements of children, and hence links family background to inequality of opportunity and intergenerational mobility. Sociologists have dominated discussions of intergenerational mobility, but in recent years economists have emphasized that the relation between the occupations, earnings, and wealths of parents and children depends on decisions by parents to spend time, money, and energy on children. Economists have used the concepts of investment in human capital and bequests of nonhuman wealth to model the transmission of earnings and wealth from parents

and children (see e.g. Conlisk 1974; Loury 1981; Becker and Tomes 1986). These models show that the relation between say the earnings of parents and children depends not only on biological and cultural endowments 'inherited' from parents, but also on the interaction between these endowments, government expenditures on children, and investments by parents in the education and other human capital of their children.

Altruism in the Family

I have followed the agnostic attitude of economists to the formation of preferences, and have not specified how quality of children is measured. One analytically tractable and plausible assumption is that parents are altruistic toward their children. By 'altruistic' is meant that the utility of parents depends on the utility of children, as in

$$U_p = U\left(z_p, U_1, \ldots, U_n\right), \tag{11}$$

where z is the consumption of parents, and U_i, $i = 1,\ldots, n$ is the utility of the ith child.

Economists have generally explained market transactions with the assumption that individuals are selfish. In Smith's famous words,

> It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk of our own necessities but of their advantages.

The assumption of selfishness in market transactions has been very powerful, but will not do when trying to understand families. Indeed, the main characteristic that distinguishes family households from firms and other organizations is that allocations within families are largely determined by altruism and related obligations, whereas allocations within firms are largely determined by implicit or explicit contracts. Since families compete with governments for control over resources, totalitarian governments have often reached for the loyalties of their subjects by attacking family traditions and the strong loyalties within families.

The preference for *own* children mentioned earlier suggests special feelings toward one's children. Sacrifices by parents to help children, and vice versa, and the love that frequently binds husbands and wives to each other, are indicative of the highly personal relations within families that are not common in other organizations (see also Ben-Porath 1980; Pollak 1985).

Although altruism is a major integrating force within families, the systematic analysis of altruism is recent, and many of its effects have not yet been determined. One significant result has been called (perhaps infelicitously) the Rotten Kid theorem, and explains the coordination of decisions among members when altruism is limited. In particular, if one member of a family were sufficiently altruistic toward other members to spend time or money on each of them, they would have an incentive to consider the welfare of the family as a whole, *even when they are completely selfish*.

The proof of this theorem is simplest when the utility of an altruist (called the 'head') depends on the combined resources of all family members. Consider a single good (x) consumed by all members: the head and n beneficiaries (not only children but possibly also a spouse and other relatives). The head's utility function can be written as

$$U_h = U(x_h, x_1, \ldots, x_n). \tag{12}$$

The budget equation would be

$$x_h + \sum_{i=h}^{n} g_i = I_h, \tag{13}$$

where I_h is the head's income, g_i is the gift to the ith beneficiary, and the price of x is set at unity. With no transactions costs, each dollar contributed would be received by a beneficiary, so that

$$x_i = I_i + g_i, \tag{14}$$

where I_i is the income of the ith beneficiary. By substitution into equation (13),

$$x_h + \sum x_i = I_h + \sum I_i = S_h. \tag{15}$$

The head can then be said to maximize the utility in (12), subject to family income (S_h).

To illustrate the theorem, consider a parent who is altruistic toward her two children, Tom and Jane, and spends say $200 on each. Suppose Tom can take an action that benefits him by $50, but would harm Jane by $100. A selfish Tom would appear to take that action if his responsibility for the changed circumstances of Jane were to go undetected (and hence not punished). However, the head's utility would be reduced by Tom's action because family income would be reduced by $50. If altruism is a 'superior good', the head will reduce the utility of each beneficiary when her own utility is reduced. Therefore, should Tom take this action, she would reduce her gift to him from $200 to less than $150, and raise her gift to Jane to less than $300. As a result, Tom would be made worse off by his actions.

Consequently, a selfish Tom who anticipates correctly the response from his parent will not take this action, even though the parent may not be trying to 'punish' Tom because she may not know that Tom is the source of the loss to Jane and the gain to herself. This theorem requires only that the head know the outcomes for both Tom and Jane and has the 'last word' (this term is due to Hirshleifer 1977).

The head has the 'last word' when gifts depend (perhaps only indirectly) on the actions of beneficiaries. In particular, if gifts to the *i*th beneficiary depend both on his income and on family income, as in

$$g_i = \psi_i(S_h) - I_i, \text{with } \frac{d\psi_i}{dS_h} > 0 \qquad (16)$$

then by substitution into equation (14),

$$x_i = I_i + g_i = \psi_i(S_h). \qquad (17)$$

The head would then have the 'last word' because x_i would be maximized by maximizing S_h; for further discussion of the Rotten Kid theorem, see Becker (1981, ch. 5), Hirshleifer (1977), and Pollak (1985).

Although this theorem is applicable even when beneficiaries are envious of each other or of the head, it does not rule out conflict in families with altruistic heads. Sibling rivalry, for example, is to be expected when children are selfish because they each want larger gifts from the head, and each would try to convince the head of his or her merits. Conflict also arises when several members are altruistic to the same beneficiaries, but not to each other. For example, if parents are altruistic to their children but not to each other, each benefits when the other spends more on the children. Married parents might readily work out an agreement to share the burden, but divorced parents have more serious conflict. Noncustodial parents (usually fathers) fall behind in their child support payments partly to shift the burden of support to custodial parents (see the discussion in Weiss and Willis 1985).

Altruism provides many other insights into the behaviour of families. For example, an efficient division of labour is possible in altruistic families without the usual principal-agent conflict because selfish as well as altruistic members consider the interests of other members. Or contrary to some opinion, bequests and gifts to children are not perfect substitutes even in altruistic families. Bequests not only transfer resources to children but also give parents the last word, which induces children to take account of the interests of elderly parents (see Becker 1981, ch. 5; and also Bernheim et al. 1986). Moreover, if public debt or social security were financed by taxes on succeeding generations that are anticipated by altruistic parents who make bequests, they would raise their bequests to offset the higher taxes paid by their children. Such compensatory reactions negate the effect of debt or social security on consumption and savings (see the detailed analysis in Barro 1974).

The Sexual Division of Labour

A sharp division of labour in the tasks performed by men and women is found in essentially all societies. Women have had primary responsibility

for child care, and men have had primary responsibility for hunting and military activity; even when both men and women engaged in agriculture, trade, or other market activities, they generally performed different tasks (see the discussion in Boserup 1970).

Substantial division of labour is to be expected in families, not only because altruism reduces incentive to shirk and cheat (see section III), but also because of increasing returns from investments in specific human capital, such as skills that are especially useful in child rearing or in market activities. Specific human capital induces specialization because investment costs are partially (or entirely) independent of the time spent using the capital. For example, a person would receive a higher return on his medical training when he puts more time into the practice of medicine. Similarly, a family is more efficient when members devote their 'working' time to different activities, and each invests mainly in the capital specific to his or her activities (see Becker 1981, 1985; for developments of this argument outside families, see Rosen 1981).

The advantages of a division of labour within families do not alone imply that women do the child rearing and other household tasks. However, the gain from specialized investments implies the traditional sexual division of labour if women have a comparative advantage in childbearing and child rearing, or if women suffer discrimination in market activities. Indeed, since a sexual division of labour segregates the activities of men and women, and since segregation is an effective way to avoid discrimination (see Becker 1981), even small differences in comparative advantage, or a small amount of discrimination against women, can induce a sharp division of labour.

Until recently, the sexual division of labour in Western countries was extreme; for example, in 1890, less than five per cent of married women in the United States were in the labour force. In 1981, by contrast, over 50 per cent even of married women with children under six were in the labour force (see Smith and Ward 1985). However, the occupations of employed men and women are still quite different, and women still do most of the child rearing and other household chores (see *Journal of Labor Economics,* January 1985).

The large growth in the labour force participation of married women during the 20th century is mainly explained by the economic development that transformed Western economies. Substitution toward market work was induced by the rise in the potential earnings of women (see Mincer 1962). Moreover, the growth in clerical jobs and in the services sector generally, gave women more flexibility in combining market work and child rearing (see Goldin 1983). In addition, the large decline in fertility during this period (see section I) greatly facilitated increased labour force participation by married women. The converse is also true, however, because the rise in participation of women discouraged child-bearing.

Divorce

Since women specialize in child care, they have been economically vulnerable to divorce and the death of their mates. All societies recognized this vulnerability by requiring long term contracts, called 'marriage', between men and women legally engaged in reproduction. In Christian societies, these contracts often could not be broken except by adultery, abandonment or death. In Islam and Asia they could be broken for other reasons as well, but husbands were required to pay compensation to their wives when they divorced without cause.

The growth of divorce during this century in Western countries has been remarkable. Essentially no divorces were granted in England prior to the 1850s (see Hollingsworth 1965), whereas now almost 30 per cent of marriages there will terminate by divorce, and the fraction is even larger in the United States, Sweden and some other Western countries (see US Bureau of the Census 1977). What accounts for this huge growth in divorce over a relatively short period of time?

The utility-maximizing rational choice perspective implies that a person wants to divorce if the utility expected from remaining married is below the utility expected from divorce, where

the latter is affected by the prospects for remarriage; indeed, most persons divorcing in Western countries now do remarry eventually (see e.g. Becker et al. 1977). This simple criterion is not entirely tautological because several determinants of the gain from remaining married can be evaluated.

Some persons become disappointed because their mates turn out to be less desirable than originally anticipated. That new information is an important source of divorce is suggested by the large fraction occurring during the first few years of marriage. Although disappointment is likely to be involved in most divorces, the large growth in divorce rates, especially the acceleration during the last 20 years, is not to be explained by any sudden deterioration in the quality of information. Instead, we look to forces that reduced the advantages from remaining in an imperfect marriage.

The strong decline in fertility over time discouraged divorce because the advantages from staying married are greater when young children are present. Conversely, fertility declined partly because divorce became more likely since married couples are less likely to have children when they anticipate a divorce (see Becker, Landes and Michael, 1977, for supporting evidence). The rise in the labour force participation of married women also lowered the gain from remaining married because the sexual division of labour was reduced, and women became more independent financially. At the same time, the labour force participation of married women increased when divorce became more likely since married women want to acquire skills that would raise their incomes if they must support themselves after a divorce.

Legislation certainly eased the legal obstacles to divorce, but empirical investigations have not found significant permanent effects on the divorce rate (see e.g. Peters 1983). Moreover, economic analysis suggests that even no-fault divorce and other radical changes in divorce legislation would not significantly affect the rate of divorce because bargaining between husbands and wives about the terms of staying married or divorcing offsets even sharp changes in divorce laws.

To show this, let income be I_h^d and I_w^d w respectively, if h and w decide to divorce, and I_h^m and I_w^m respectively, if they remain married. The budget equation is

$$x_h^d + x_w^d = I_h^d + I_w^d = I^d \tag{18}$$

when divorced, and

$$x_h^d + x_w^m = I_h^m + I_w^m = I^m \tag{19}$$

when married. I suggest that the decision to divorce is largely independent of divorce laws, and depends basically on whether $< \text{dout} > I^{\,d}$ $I^{\,m}$, because both h and w can be made better off by divorce when $I^{\mathrm{d}} > I^{\mathrm{m}}$ and by remaining married when $I^{\mathrm{m}} > I^{\mathrm{d}}$.

Consider, for example, a comparison between unilateral or no-fault divorce, and divorce only by mutual consent. Assume that the husband appears to gain from divorce $\left(I_h^d > I_h^m\right)$ but the apparent loss to the wife is greater, so that $I^{\mathrm{d}} < I^{\mathrm{m}}$. If divorce were unilateral, he might be tempted to seek a divorce even when she would be greatly harmed. However, she could change his mind by offering a bride (b_h) that would make both of them better off by staying married:

$$x_h^m + x_w^m > I_h^d, \text{ and } I_w^m - b_h > I^d \tag{20}$$

This bribe is feasible because $x_h^m + x_w^m = I^m > I^d$. He would then prefer to remain married, even if he could divorce without her consent. Note that they would also decide to remain married if divorce required mutual consent because at least one of them must be made worse off by divorce.

Divorce rates have been affected less by legislation that has regulated the conditions for divorce than by legislation that has affected the gains from divorce. For example, aid to mothers with dependent children and negative income taxes encourage divorce by providing poorer women with child support and 'alimony' (see Hannan et al. 1977).

Marriage

Marriages can be said to take place in a 'market' that 'assigns' men and women to each other or to

remain single until better opportunities come along. An optimal assignment in an efficient market with utility-maximizing participants has the property that persons not assigned to each other could not be made better off by marrying each other.

In all societies, couples tend to be of similar family background and religion, and are positively sorted by education, height, age, and many other variables. The theory of assignments in efficient markets explains positive assortative mating by complementarity, or 'superadditivity', in household production between the traits of husbands and wives. Efficient assignments also partly explain altruism between husbands and wives: persons 'in love' are likely to marry because, at the detached level of formal analysis, love can be considered one source of 'complementarity'.

Associated with optimal assignments are imputations that determine the division of incomes or utilities in each marriage. Equilibrium incomes have the property that

$$I_{ii}^{m} + I_{ii}^{f} = I_{ii}, \tag{21}$$

and

$$I_{ii}^{m} + I_{ii}^{f} \geq I_{ii}, \quad i \neq j \tag{22}$$

where I_{ij} the output from a marriage of the ith man (m_i) to the j the women (f_j), and I_{ii}^{m} and I_{ii}^{f} are the incomes of m_i and f_j, respectively. The inequality in equation (22) indicates the $\{ii\}$ is an optimal assignment because m_i and f_j $j \neq i$, could not be made better off by marrying each other instead of their assigned mates (f_i and m_j, respectively). Equilibrium incomes include dowries, bride prices, leisure and 'power' (further discussion can be found in Becker 1974, 1981; the analysis of optimal assignments in Gale and Shapley 1962; and Roth 1984, is less relevant to marriage because equilibrium prices – i.e. incomes – are not considered).

Many of the forces in recent decades that reduce the gain from remaining married (see section V) have also raised the gain from delaying first marriage and remarriage. These include the decline in fertility and the rise in labour force participation of married women. The reduced incentive to marry in Western societies is evident from the rapid increase in the number of couples living together without marriage, and in the number of births to unmarried women. Nevertheless, even in Scandinavia, where the trend toward cohabitation without marriage has probably gone furthest, married persons are still far more likely to remain together and to produce children than are persons who cohabit without marriage (for Swedish evidence, see Trost 1975).

F

Summary and Concluding Remarks

Families are important producers as well as spenders. Their primary role has been to supply future generations by producing and caring for children, although they also help protect members against ill health, old age, unemployment, and other hazards of life.

Families have relied on altruism, loyalty, and norms to carry out these tasks rather than the contracts found in firms. Altruism and loyalty are concepts that have not been utilized extensively to analyse market transactions, and our understanding of their implications is only beginning. Yet a much more complete understanding is essential before the behaviour and evolution of families can be fully analysed.

Firms and families compete to organize the production and distribution of goods and services, and activities have passed from one to the other as scale economies, principal-agent problems, and other forces dictated. Agriculture and many retailing activities have been dominated by family firms that combine production for the market with production for members. Presumably, such hybrid organizations are important when altruism and loyalty are more effective than contracts in organizing market production (see Becker 1981, ch. 8; Pollak 1985), and when the production and care of children complements production for the market.

Families in Western countries have changed drastically during the past thirty years; fertility declined below replacement levels, the labour

force participation of married women and divorce soared, cohabitation and births to unmarried women became common, many households are now headed by unmarried women with dependent children, a large fraction of the elderly either live alone or in nursing homes, and children from first and second, sometimes even third, marriages frequently share the same household.

Nevertheless, obituaries for the family are decidedly premature. Families are still crucial to the production and rearing of children, and remain important protectors of members against ill-health, unemployment, and many other hazards. Although the role of families will evolve further in the future, I am confident that families will continue to have primary responsibility for children, and that altruism and loyalty will continue to bind parents and children.

See Also

- ▶ Altruism
- ▶ Family Planning
- ▶ Fertility
- ▶ Gender
- ▶ Household Production
- ▶ Human Capital
- ▶ Inequality Between the Sexes
- ▶ Value of Time
- ▶ Women's Wages

Bibliography

Barro, R.J. 1974. Are government bonds net wealth? *Journal of Political Economy* 82(6): 1095–1117.

Becker, G.S. 1974. A theory of marriage: Part II. *Journal of Political Economy* 82(2): S11–S26, part II.

Becker, G.S. 1981. *A Treatise on the family*. Cambridge, MA: Harvard University Press.

Becker, G.S. 1985. Human capital, effort, and the sexual division of labor. *Journal of Labor Economics* 3(1): 533–558, Part II.

Becker, G.S., and N. Tomes. 1986. Human capital and the rise and fall of families. *Journal of Labor Economics* 4(2, pt. 2): S1–S39.

Becker, G.S., E.M. Landes, and R.T. Michael. 1977. An economic analysis of marital instability. *Journal of Political Economy* 85(6): 1141–1187.

Ben-Porath, Y. 1980. The F-connection: Families, friends, and firms and the organization of exchange. *Population and Development Review* 6(1): 1–30.

Bernheim, B.I., A. Schleiffer, and L.H. Summers. 1986. Bequests as a means of payment. *Journal of Labor Economics* 4(3): S151–S182, pt. 2.

Blake, J. 1981. Family size and the quality of children. *Demography* 18(4): 421–442.

Boserup, E. 1970. *Woman's role in economic development*. London: Allen & Unwin.

Conlisk, J. 1974. Can equalization of opportunity reduce social mobility? *American Economic Review* 64(1): 80–90.

De Tray, D.N. 1973. Child quality and the demand for children. *Journal of Political Economy* 81(2): S70–S95, Pt II.

Gale, D., and L.S. Shapley. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69(1): 9–15.

Gardner, B. 1973. Economics of the size of North Carolina rural families. *Journal of Political Economy* 81(2): S99–S122, Part II.

Goldin, C. 1983. The changing economic role of women: A quantitative approach. *Journal of Interdisciplinary History* 13(4): 707–733.

Hannan, M.T., N.B. Tuma, and L.P. Groeneveld. 1977. Income and martial events: Evidence from an income maintenance experiment. *American Journal of Sociology* 82(6): 611–633.

Hirshleifer, J. 1977. Shakespeare vs. Becker on altruism: The importance of having the last word. *Journal of Economic Literature* 15(2): 500–502.

Hollingsworth, T.H. 1965. The demography of the British peerage. Supplement to *Population Studies* 18(2).

Jaffe, A.J. 1940. Differential fertility in the white population in early America. *Journal of Heredity* 31(9): 407–411.

Locay, L. 1987. *Population density of the North American Indians*. Cambridge, MA: Harvard University Press.

Loury, G.C. 1981. Intergenerational transfers and the distribution of earnings. *Econometrica* 49(4): 843–867.

Malthus, T.R. 1798. *An essay on the principle of population*. Reprinted. London: J.M. Dent, 1958.

Marshall, A. 1890. *Principles of economics*. London: Macmillan.

Mill, J.S. 1848. *Principles of political economy, with some of their applications to social philosophy*. Reprinted. New York: Colonial Press, 1899.

Mincer, J. 1962. Labor force participation of married women. In *Aspects of labor economics*. Princeton: Princeton University Press.

Peters, E. 1983. The impact of state divorce laws on the marital contract: Marriage, divorce, and marital property settlements. Discussion Paper No. 83–19. Economics Research Center/NORC.

Pollak, R.A. 1985. A transactions cost approach to families and households. *Journal of Economic Literature* 23(2): 581–608.

Rosen, S. 1981. Specialization and human capital. *Journal of Labor Economics* 1(1): 43–49.
Roth, A. 1984. The evolution of the labor market for medical interns and residents: A case study in game theory. *Journal of Political Economy* 92(6): 991–1016.
Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. Reprinted. New York: Modern Library, 1937.
Smith, J.P., and M.P. Ward. 1985. Time series growth in the female labor force. *Journal of Labor Economics* 3(1): 559–590, Part II.
Trost, J. 1975. Married and unmarried cohabitation: The case of Sweden and some comparisons. *Journal of Marriage and the Family* 37(3): 677–682.
US Bureau of the Census, 1977. *Current population reports*. Series P–20, No. 308, Fertility of American Women: June, 1976.
Weiss, Y., and R. Willis. 1985. Children as collective goods and divorce settlements. *Journal of Labor Economics* 3(3): 268–292.
Wrigley, E.A., and R.S. Schofield. 1981. *The population history of England 1541–1871*. Cambridge, MA: Harvard University Press.

Family Decision Making

Shelly Lundberg and Robert A. Pollak

Abstract

The classic unitary model assumes that households maximize a household utility function and implies resource 'pooling' – household behaviour does not depend on individuals' control over resources within the household. Since the 1980s, economists have modified the unitary model in ways that have theoretical, empirical and practical implications. Non-unitary alternatives based on joint decision-making by individual family members with distinct preferences broaden the range of observable behaviour consistent with economic rationality. Many non-unitary models imply that both individuals' control over resources and 'environmental factors' can affect intra-household allocation. Empirical evidence has consistently rejected income pooling and, hence, the unitary model.

Keywords

Altruist model of the family; Bargaining; Collective model of the household; Consensus model of the family; Cooperative bargaining model of marriage; Cooperative games; Cournot–Nash equilibrium; Family decision-making; Gender specialization; Household production and public goods; Marriage and divorce; Non-cooperative bargaining model of marriage; Rotten kid theorem; Self-enforcing agreements; Separate spheres bargaining model; Slutsky symmetry; Two-stage games; Unitary and non-unitary models of the household

JEL Classifications

B4

Economic models of consumer demand and labour supply begin with an individual economic agent choosing actions that maximize his or her utility subject to a budget constraint. How can we reconcile this individualistic theory of the consumer with the reality that people tend to live, eat, work and play in families? Economists have dealt with a possible multiplicity of decision-makers in the family in two ways. The first, in ascendancy until the 1980s, was the unitary approach – treating the family as though it were a single decision-making agent, with a single pooled budget constraint and a single utility function that includes the consumption and leisure time of every family member. The second approach, pioneered in the early 1980s by Manser and Brown and by McElroy and Horney, was to model family behaviour as the solution to a cooperative bargaining game. Other non-unitary approaches have subsequently been developed, including the 'collective' model of Chiappori, extensions of the cooperative models of Manser–Brown and McElroy–Horney, and various non-cooperative models.

Most non-unitary models of family behaviour allow two decision makers – the husband and the wife; children are customarily excluded from the

set of decision-making agents, though they may be recognized as consumers of goods chosen and provided by loving or dutiful parents. Bargaining models have also been used to analyse interactions between parents and adolescent or young adult children, and between elderly parents and adult children. These interactions may involve family members living in different households, and, in many of these models, who lives with whom is endogenous. As a class, non-unitary models are consistent with a wider range of behaviour than unitary models. The empirical implications of specific non-unitary models of the family depend upon their assumptions about preferences, opportunities, and the form of the game.

Unitary Models

Two models provide the theoretical underpinning of the unitary, or common preference, approach to family behaviour: Samuelson's (1956) consensus model and Becker's (1974, 1981) altruist model. The consensus model was introduced by Samuelson to exhibit the conditions under which family behaviour can be rationalized as the outcome of maximizing a single utility function. Consider a two-member family consisting of a husband and a wife. Each partner has an individual utility function that depends on his or her private consumption of goods, but, by consensus, they agree to maximize a social welfare function incorporating their individual utilities, subject to a joint budget constraint that pools the income received by the two spouses. Then we can analyse the household's observed aggregate expenditure pattern as though the family were a single agent maximizing a utility function (that is, the consensus social welfare function). That is, the household maximizes $U(c^h, c^w)$, where c^h and c^w are the private consumptions of husband (h) and wife (w), subject to the budget constraint $p(c^h + c^w) = y = y^h + y^w$ which pools the individual incomes of husband and wife. This problem generates demand functions $c^i = f^i(p, y)$ that depend only on prices and total family income and that have standard properties provided the utility functions are well behaved. Thus, the comparative statics of traditional consumer demand theory apply directly to family behaviour under the consensus model. Samuelson did not, however, purport to explain how the family achieves a consensus regarding the joint welfare function, or how this consensus is maintained.

Becker's (1974, 1981) altruist model addresses these questions, and also provides an account of how resources are distributed within the family. In Becker's model, the family consists of a group of purely selfish but rational 'kids' and one altruistic parent whose utility function reflects his concern for the well-being of other family members. Becker argues that the presence of an altruistic parent who makes positive transfers to each member of the family is sufficient to induce the selfish kids to act in an apparently unselfish way. The altruistic parent will adjust transfers so that each 'rotten kid' finds it in his interest to choose actions that maximize family income. The resulting distribution is the one that maximizes the altruist's utility function subject to the family's resource constraint, so the implications of the altruist model for family demands coincide with those of the consensus model (see Bergstrom 1989 for a discussion of the conditions under which the rotten kid theorem holds and does not hold).

Unitary models provide a simple, powerful mechanism for generating demand functions and establishing their comparative statics for use in applied problems. Since the introduction of the bargaining paradigm however, these models have been criticized on both empirical and theoretical grounds. We first discuss the theoretical criticisms, and then turn to the accumulating empirical evidence inconsistent with the unitary model.

Dissatisfaction with unitary models on theoretical grounds has been the product of serious study of marriage and divorce. Models of marriage and divorce require a theoretical framework in which agents compare their expected utilities inside marriage with their expected utilities outside marriage, but the individual utilities of husband and wife outside marriage cannot be recovered from the social welfare function that generates consumption, labour supply, fertility, and other behaviour within marriage. If the analysis of

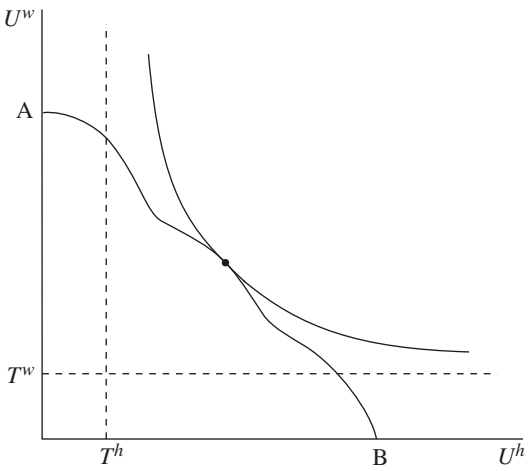
marriage and divorce is awkward, the analysis of marital decisions in the shadow of divorce is even more so. If unilateral divorce is possible, individual rationality implies that marital decisions cannot leave either husband or wife worse off than they would be outside the marriage. This individual rationality requirement, however, alters the comparative statics of the model, and destroys the correspondence between the behaviour of a single utility maximizing agent and the behaviour of a family.

Non-unitary Models

Cooperative Bargaining Models

A viable alternative to unitary models of the family must recognize, in a non-trivial fashion, the involvement of two or more agents in determining family consumption. Bargaining models from cooperative game theory, first applied to marriage by Manser and Brown (1980) and by McElroy and Horney (1981), satisfy these conditions. A typical cooperative bargaining model of marriage begins with a family that consists of only two members, a husband and a wife. Each has a utility function that depends on his or her consumption of private goods ($U^h(c^h)$ for the husband and $U^w(c^w)$ for the wife). If agreement is not reached, then the payoff received is represented by the 'threat point', ($T^h(Z)$, $T^w(Z)$) – the utilities associated with a default outcome of divorce or, in the 'separate spheres' model of Lundberg and Pollak (1993), a non-cooperative equilibrium within the marriage. The threat point depends, in turn, upon a set of exogenous distribution factors Z that influence individual well-being in the default outcome.

The Nash bargaining model provides the leading solution concept in bargaining models of marriage. Nash bargaining implies that the couple maximizes the Nash product function $N = [U^h(c^h) - T^h(Z)]\,[U^w(c^w) - T^w(Z)]$ subject to a pooled budget constraint, and this results in demand functions of the form $c^i = f^i(p, y, Z)$. Thus demands and individual utilities depend upon the distribution factors Z, which may include individual incomes y^h and y^w. This solution can be illustrated by a diagram in utility space (Fig. 1),



Family Decision Making, Fig. 1 The Nash bargaining solution

where AB is the utility-possibility frontier. Nash (1950) shows that a set of four axioms, including Pareto efficiency – which ensures that the solution lies on the utility-possibility frontier – uniquely characterize the Nash bargaining solution.

The utility received by husband or wife in the Nash bargaining solution depends upon the threat point: the higher one's utility at the threat point, the higher one's utility in the Nash bargaining solution. This dependence is the critical empirical implication of Nash bargaining models: family demands depend, not only on prices and total family income, but also on determinants of the threat point.

In divorce-threat bargaining models, the threat point is the maximal level of utility attainable outside the marriage. Hence, the threat point depends on wage rates and on the assets each spouse would take if the marriage were to end in divorce. The divorce threat point is also likely to depend on environmental factors (extra-household environmental parameters, or EEPs, in McElroy's 1990, terminology) that do not directly affect marital utility, such as conditions in the remarriage market and the income available to divorced men and women. The family demands that result from divorce-threat marital bargaining will therefore depend upon these parameters as well.

In the separate spheres bargaining model of Lundberg and Pollak (1993), the threat point is internal to the marriage, not external as in divorce-threat bargaining models. The husband and wife settle their differences by Nash bargaining, but the alternative to agreement is an inefficient non-cooperative equilibrium within marriage. In this non-cooperative equilibrium, each spouse voluntarily provides household public goods, choosing actions that are utility-maximizing, given the actions of their partner. Divorce may be the ultimate threat available to marital partners in disagreement, but a non-cooperative marriage in which the spouses receive some benefits due to joint consumption of public goods may be a more plausible threat in day-to-day marital bargaining.

The introduction of this internal threat point has important implications, because the separate spheres model generates family demands that, under some circumstances, depend not on who receives income after divorce, but on who receives (or controls) income within the marriage. Lundberg and Pollak assume gender specialization in the non-cooperative provision of household public goods, with the husband providing one good out of his own resources, and the wife providing a separate good from her individual resources. This specialization occurs because socially prescribed gender roles provide a focal point for non-cooperative bargaining. The individual reaction functions in this game determine a Cournot–Nash equilibrium in which the public goods contributions may be inefficiently low, and may depend upon the distribution of individual incomes within the family.

As the divorce-threat and separate spheres models show, cooperative bargaining does not necessarily imply income pooling, that is, the property that demands depend only on total household income, rather than its separate components. Bargained outcomes depend upon the threat point, and the income controlled by husband and wife will affect family behaviour (and the relative well-being of men and women within marriage) if this control influences the threat point. This dependence implies that public policy (for example, taxes and transfers) need not be neutral in their effects on distribution within the family. Also, the absence of pooling and the presence of extra-household environmental parameters in family demands yield a model that can be tested against the unitary alternative. For example, changes in the welfare payments available to divorced mothers, or in the laws defining marital property and regulating its division upon divorce, should affect distribution between men and women in two-parent families through their effect on the threat point.

The 'Collective' Approach

Most models of the family either assume or conclude that family behaviour is Pareto efficient. Unitary models ensure Pareto efficiency by assuming a family social welfare function that is increasing in the utilities of all family members: when such a utility function is maximized, no member can be made better off without making another worse off. Cooperative bargaining models characterize the equilibrium distribution by means of a set of axioms, one of which is Pareto efficiency.

Pareto efficiency is the defining property of the 'collective model' of Chiappori (1988, 1992). Rather than applying a particular cooperative or non-cooperative bargaining model to the household allocation process, Chiappori assumes only that equilibrium allocations are Pareto efficient. He demonstrates that, given a set of assumptions including weak separability of public goods and the private consumption of each family member, Pareto efficiency implies, and is implied by, the existence of a 'sharing rule'. Under a sharing rule, the family acts as though decisions were made in two stages: first total family income is divided between public goods and the private expenditures of each individual, and then each individual allocates his or her share among private goods. The collective model implies a set of testable restrictions on the response of household demands to 'distribution factors' that affect the household's sharing rule.

Non-cooperative Bargaining Models

The use of models that assume Pareto efficiency of outcomes relies on the judgement that information within families is relatively good (or at least not asymmetric) and that members are able to make binding, costlessly enforceable agreements.

Since legal institutions do not provide for external enforcement of contracts regarding consumption, labour supply, and allocation within marriage, however, the binding-agreement assumption is unappealing.

Non-cooperative game theory focuses on self-enforcing agreements. It is possible for non-cooperative bargaining to yield Pareto efficient outcomes under certain conditions. For example, repeated non-cooperative games have multiple equilibria which are sustained by credible threats of punishment, and some of these equilibria are Pareto efficient. One of the benefits of modelling distribution within marriage as a non-cooperative game is the opportunity to treat efficiency as endogenous, potentially dependent upon the institutions and social context of marriage in a particular society and upon the characteristics of the marital partners.

The prevalence of destructive or wasteful phenomena such as domestic violence and child abuse, as well as the demand for marriage counselling and family therapy, suggests that we consider the possibility that family behaviour is sometimes inefficient. Other researchers have pointed to gender segmentation in the management of businesses or agricultural plots in many countries as evidence of an essentially non-cooperative, and possibly inefficient, family environment. One piece of evidence is provided by Udry (1996), who finds that in Burkina Faso the marginal product of land controlled by women is below the marginal product of land controlled by men and concludes that the household allocation of inputs to male- and female-controlled agricultural plots is inefficient.

Intertemporal Models

In dynamic bargaining models, decisions made in one period can alter the relative bargaining power of individual family members in future periods. If family members cannot agree on rules for sharing household resources in the future, and make credible promises to obey such rules, then inefficiencies of the standard 'hold-up' variety will result. Lundberg and Pollak (2003) model the two-earner couple location problem as a two-stage game in which a couple must decide where to live and whether to stay together without being able to make binding commitments about allocation in the new location. Lundberg and Pollak show that the equilibrium of this two-stage game need not be efficient even if the second-stage game is conditionally efficient (that is, efficient given the location determined at the first stage).

Even if prospective spouses can make binding agreements in the marriage market, they cannot make agreements with potential spouses they have not yet met. Konrad and Lommerud (2000) show that individuals will over-invest in education prior to marriage to increase their marital bargaining power, even if they expect to bargain cooperatively once they find and marry a spouse. Models of limited commitment in marriage can also be applied to decisions about childbearing, career choice and work effort.

Empirical Evidence

Recent empirical evidence suggests that the restrictions imposed on demand functions by unitary models are not well supported. Rejections of the family income pooling hypothesis, in particular, have been most influential in weakening economists' attachment to unitary models. Unitary models imply that the fraction of income received or controlled by one family member should not influence demands, given total family income. A large number of recent empirical studies have rejected pooling, finding that earned and unearned income received by the husband or wife significantly affect demand patterns when total income or expenditure is held constant. Some studies find that children appear to do better when their mothers control a larger fraction of family resources (Thomas 1990; Haddad and Hoddinott 1994). These results are inconsistent with the unitary framework, but consistent with both bargaining models (provided individual incomes affect the threat point) and with the collective model (provided individual incomes are included among the 'distribution factors' that influence the household's sharing rule).

The collective model imposes, in addition, a proportionality restriction on the influence of

distribution factors on demands. The ratio of the marginal propensities to consume any two goods must be the same for all sources of income, for example, because individual incomes affect consumption only through the sharing rule. A generalization of Slutsky symmetry in price effects can also be derived (Browning and Chiappori 1998). A series of empirical tests have found that consumption expenditures in households reject the unitary framework but are generally consistent with the collective model (for example, Bourguignon et al. 1993; Browning and Chiappori 1998).

Tests of the unitary model against non-unitary alternatives require a measure of husband's and wife's relative control over resources. Relative earnings would seem to be an attractive candidate for this measure, since labour income is by far the largest component of family income, and earnings data are readily available and reliably measured. Also, the earnings of wives relative to husbands have increased dramatically in the United States and many other countries, and we would like to assess the distributional consequences, if any, of this change. The difficulty with this approach is that earnings are clearly endogenous with respect to household time-allocation decisions. Earnings are the product of hours worked, a choice variable, and hourly wage rates, which measure the prices of time for husband and wife and therefore enter demand functions directly in the unitary model. This implies that households with different ratios of wife's earnings to husband's earnings are likely to face different prices and may have different preferences.

One might try to avoid these problems by testing the pooling of unearned income rather than earnings. Unearned income is not contaminated by price effects, but most unearned income sources are not entirely exogenous with respect to past or present household behaviour. Schultz (1990), who like Thomas (1990) uses unearned income to test the pooling hypothesis, points out that variations in unearned income over a cross-section are likely to be correlated with other (possibly unobservable) determinants of consumption. For example, property income reflects, to a considerable extent, accumulated savings and is therefore correlated with past labour supply and, if those who worked a lot in the past continue to do so, with current labour supply. Public and private transfers may be responsive to household distress due to unemployment or bad health, and may be related to expenditures through the events that prompted them. *Unexpected* transfers such as lottery winnings, unexpected gifts or unexpected bequests will affect resources controlled by individuals without affecting prices, but are likely to be sporadic and unimportant for most families.

Other standard empirical proxies for the relative bargaining power of husbands and wives (or, in the terminology of the collective model, distribution factors) include the relative ages, educations, or measures of family background of husband and wife. The interpretation of these factors, however, is contaminated by assortative mating on unobserved characteristics. It would be unwise to assume that a highly educated woman married to a man with less education has relatively more control over the allocation of household resources without controlling for other personal characteristics that affected the decision of this couple to marry in the first place. The same critique applies to measures of relative assets brought to the marriage by the husband and wife, even when they maintain separate ownership of these assets during marriage and divorce.

The ideal test of the pooling hypothesis, and therefore of the unitary family model, would be based on an experiment in which some husbands and some wives were randomly selected to receive income transfers. A less-than-ideal test could be based on a 'natural experiment' in which some family members receive an exogenous income change, and one can study a constant population of families before and after the change. Several studies exploiting such policy changes have found evidence against income pooling, and have also supported the hypothesis that women have a greater propensity, on average, to spend on children's goods.

Lundberg et al. (1997) examine the effects of a policy change in the United Kingdom that transferred a substantial child allowance from husbands to wives in the late 1970s. They find strong evidence that a shift towards relatively

greater expenditures on women's goods and children's goods coincided with this income redistribution, and interpret this as a rejection of the pooling hypothesis. Duflo (2000) studied the effect of an extension of the South African Old Age Pension on children's health and nutrition, and found that payments to grandmothers had a substantial effect on these outcomes, especially for girls, while payments to grandfathers had no effect. These results both reject a unitary framework for multi-generation families, and support the hypothesis that children benefit from female control of household resources. Tests of pooling using PROGRESA, a public cash transfer programme in Mexico directed at women, have been more complicated. A random assignment social experiment, PROGRESA had a substantial income effect and benefits were conditional on child school enrolment. Attanasio and Lechene (2002) reject household pooling using PROGRESA data, and Rubalcava et al. (2004) find that these transfers to women were more likely to be spent on child goods, improved nutrition, and investments in small livestock than other household income.

One important implication of non-unitary models of the household is that government programmes targeted to particular individuals within households may affect the intra-household allocation. Even if, as rejections of the unitary model suggest, targeted transfers are effective in the short run, we cannot conclude that targeted transfers will be effective in the long run. Lundberg and Pollak (1993) show that the long-term effects of such policy changes on intra-household allocation may be very different from the short-term effects, as adjustments occur in the marriage market of subsequent cohorts. If prospective couples can make binding agreements when they marry, then the distributional effects of policy can be offset by subsequent generations of families. Even if such marital agreements are not possible, changes in the expected gains to marriage will affect who marries whom and who marries at all, and this will also affect the long-run distributional effects of policy. Cross-sectional studies of intra-household allocation that use state variation in policy or laws (such as divorce laws or property settlement rules) will be estimating the equilibrium effects of long-standing differences in policy, including any marital sorting effects.

Conclusion

The classic unitary model assumes that households maximize a household utility function subject to household resource and technology constraints. Unitary models imply income or resource 'pooling' – household behaviour does not depend on individual control over resources within the household. Since the 1980s, economists have modified the unitary model in ways that have theoretical, empirical and practical implications. Non-unitary alternatives based on joint decision-making by individual family members with distinct preferences broaden the range of observed behaviour consistent with economic rationality. Non-unitary models also permit the analysis of marriage and divorce within the same framework as household demands and the labour supply of household members. Unlike unitary models, many non-unitary models imply that both individual control over resources and 'environmental factors', such as divorce laws, that affect the well-being of individuals outside the household can affect intrahousehold allocation. Empirical evidence has consistently rejected income pooling and, hence, the unitary model.

See Also

- ▶ Collective Models of the Household
- ▶ Family Economics
- ▶ Intrahousehold Welfare
- ▶ Marriage and Divorce

Bibliography

Attanasio, O., and V. Lechene. 2002. Tests of income pooling in household decisions. *Review of Economic Dynamics* 5: 720–748.

Becker, G.S. 1974. A theory of social interactions. *Journal of Political Economy* 82: 1063–1093.

Becker, G.S. 1981. Altruism in the family and selfishness in the market place. *Economica* 48: 1–15.

Becker, G.S. 1991. *A treatise on the family*. Enlarged ed. Cambridge, MA: Harvard University Press.
Bergstrom, T. 1989. A fresh look at the rotten kid theorem and other household mysteries. *Journal of Political Economy* 97: 1138–1159.
Bourguignon, F., M. Browning, P.-A. Chiappori, and V. Lechene. 1993. Intra household allocation of consumption: A model and some evidence from French data. *Annales d'Economie et de Statistique* 29: 138–156.
Browning, M., and P.-A. Chiappori. 1998. Efficient intra-household allocations: A general characterization and empirical tests. *Econometrica* 66: 1241–1278.
Chiappori, P.-A. 1988. Rational household labor supply. *Econometrica* 56: 63–89.
Chiappori, P.-A. 1992. Collective labor supply and welfare. *Journal of Political Economy* 100: 437–467.
Duflo, E. 2000. Child health and household resources in South Africa: Evidence from the old age pension program. *American Economic Review* 90: 393–398.
Haddad, L., and J. Hoddinott. 1994. Women's income and boy–girl anthropometric status in the Côte d'Ivoire. *World Development* 22: 543–553.
Konrad, K.A., and K.E. Lommerud. 2000. The bargaining family revisited. *Canadian Journal of Economics* 33: 471–487.
Lundberg, S.J., and R.A. Pollak. 1993. Separate spheres bargaining and the marriage market. *Journal of Political Economy* 101: 988–1010.
Lundberg, S.J., and R.A. Pollak. 2003. Efficiency in marriage. *Review of Economics of the Household* 1: 153–167.
Lundberg, S.J., R.A. Pollak, and T.J. Wales. 1997. Do husbands and wives pool their resources? Evidence from the United Kingdom child benefit. *Journal of Human Resources* 32: 463–480.
Manser, M., and M. Brown. 1980. Marriage and household decision-making: A bargaining analysis. *International Economic Review* 21: 31–44.
McElroy, M.B. 1990. The empirical content of Nash-bargained household behavior. *Journal of Human Resources* 25: 559–583.
McElroy, M.B., and M.J. Horney. 1981. Nash-bargained household decisions: Toward a generalization of the theory of demand. *International Economic Review* 22: 333–349.
Nash, J. 1950. The bargaining problem. *Econometrica* 18: 155–162.
Rubalcava, L., G. Teruel, and D. Thomas. 2004. Spending, saving, and public transfers paid to women. California Center for Population Research. Online Working Paper Series. CCPR-024–04. Online. Available at http://www.ccpr.ucla.edu/ ccprwpseries/ccpr_024_04.pdf. Accessed 22 Apr 2007.
Samuelson, P.A. 1956. Social indifference curves. *Quarterly Journal of Economics* 70: 1–22.
Schultz, T.P. 1990. Testing the neoclassical model of family labor supply and fertility. *Journal of Human Resources* 25: 599–634.
Thomas, D. 1990. Intra-household resource allocation: An inferential approach. *Journal of Human Resources* 25: 635–664.
Udry, C. 1996. Gender, agricultural production and the theory of the household. *Journal of Political Economy* 104: 1010–1046.

Family Economics

John Ermisch

Abstract

Family economics is the application of the analytical methods of microeconomics to family behaviour. It aims to improve our understanding of resource allocation and the distribution of welfare within the family, investment in children and inter-generational transfers, family formation and dissolution and how families and markets interact. In family economics, non-market interactions are crucial for family behaviour and individual welfare.

Keywords

Altruism; Becker, G; Child nutrition and mortality; Collective models of the household; Demographic transition; Family decision-making; Family economics; Family planning; Fertility in developed countries; Fertility in developing countries; Human capital; Intergenerational income mobility; Intergenerational transfers; Intrahousehold welfare; Labour supply; Malthus, T; Marriage and divorce; Rotten Kid Theorem; Shadow pricing; Women's work and wages

JEL Classifications

D11

Family economics is the application of the analytical methods of microeconomics to family behaviour. It aims to improve our understanding of resource allocation and the distribution of welfare within the family, investment in children and

inter-generational transfers, family formation and dissolution, and how families and markets interact. Family economics lifts the lid on the 'black box' of the family, within which non-market interactions are crucial for family behaviour and individual welfare. It analyses how markets affect family behaviour and on how family context affects market behaviour, such as labour supply and consumer demand, thereby linking family economics with traditional fields of economics.

Medical and social sciences indicate the importance of nutritional, cognitive and emotional development during childhood for a person's lifetime health and prosperity, and these developments are a product of parents' actions, including family break-up. Acquiring a better understanding of family formation and dissolution and of decisions within the family, particularly as these are affected by elements of people's budget constraints, is an important prerequisite for understanding how public policy can influence the family.

In a broad sense, family economics has been around for over two hundred years. Thomas Malthus believed that human fertility was determined by the age at marriage and frequency of coition during marriage. He contended that an increase in people's income would encourage them to marry earlier and have sexual intercourse more often. Modern economic theories of fertility generalized the Malthusian theory (starting with Becker 1960), and Gary Becker subsequently developed a broader economic analysis of the family (Becker 1981), which forms the foundation for today's family economics (Ermisch 2003).

What Influences Family Decisions?

Individualism needs to be the foundation of family economics if we are to analyse the impacts of public policies and technological developments on the welfare of *individuals.* In particular, decisions about marriage and divorce must make comparisons between *individual* welfare within and outside a couple. The family is best viewed as a 'governance structure' for organizing its activities rather than as a preference ordering augmented by home production technology, or as a set of long-term contracts (Pollak 1985). This suggests that bargaining models, in which alternatives and 'threat points' affect intra-family allocation and distribution, provide a useful framework for analysing family behaviour. A bargaining approach naturally focuses on the structure of family membership and its internal organization (for example, comparing an intact nuclear family with divorced parents), and allows decisions to evolve in a flexible way.

A fruitful starting point is to assume that all individuals act to maximize their welfare as they evaluate it, given the predicted behaviour of others in the family. Some authors have adopted this non-cooperative approach to studying family choices (for example, Konrad and Lommerud 1995). But, in many circumstances (for example, the co-resident family), cooperative behaviour is a better representation of family behaviour because of repeated interaction between family members, which facilitates information flows and monitoring. Nevertheless, family members must obtain welfare from cooperation that is at least as great as they would achieve from a non-cooperative outcome, although in some circumstances divorce may be a credible threat affecting decisions within the family (Bergstrom 1996).

Cooperation achieves an efficient allocation of resources within the family. Individual welfare depends, in general, on individual incomes and prices and possibly other 'distribution factors' like marriage market conditions, divorce laws and other institutions (Browning and Chiappori 1998), whose influence reflects bargaining between family members. For example, an increase in the mother's income may have two effects on family choices. It increases family resources, expanding welfare-enhancing opportunities for all family members. It also may increase her threat point (bargaining power), which pushes family choices in her favour, thereby increasing her welfare relative to the father's. Put differently, her income affects the position of the family's utility possibility frontier and also the position on it. If, for example, mothers' preferences put more weight on children than fathers' preferences do, then an increase in her share of family income

would increase expenditure on children. If this is the case, then children do better when mothers control more of family resources, developments which improve women's earning opportunities affect the distribution of welfare within families, and it is possible to target policies on individuals within families. In a dynamic setting, in which current decisions affect future bargaining power, efficiency is harder to sustain because of the difficulty of making binding commitments, but individual incomes and distribution factors still affect intra-family allocation.

One particular decision-making rule has been an important part of family economics: the family maximizes the welfare of an 'effective altruist'. It is in fact a special case of the cooperative (efficient outcomes) framework just discussed. A person is said to be *altruistic* toward someone if his or her welfare depends on the welfare of that person. Altruism is usually defined more narrowly, by what have been called 'caring' preferences: the altruist's 'social' utility takes the form $W^A[U^A(\boldsymbol{x}_A,\boldsymbol{G}),\ U^B(\boldsymbol{x}_B,\boldsymbol{G})]$, where $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ are vectors of private goods consumed by persons A and B respectively, $\boldsymbol{G}$ is a vector of public goods and $U^A(\cdot)$ and $U^B(\cdot)$ are 'private' utility indices for each person. The altruist A does not care how (in terms of $\boldsymbol{x_B}$ and $\boldsymbol{G}$) a given level of private utility is obtained by his/her beneficiary B.

Caring preferences limit only the relevant range of the utility possibility frontier expressed in terms of *private* preferences. Some family decision rule is still needed to determine the point on the frontier that is chosen, but there are circumstances in which caring preferences can produce distinctive predictions. Suppose that a wife and her husband care for each other, and her share of joint income is sufficiently large that she is making transfers to him to ensure that his welfare is not too low. To use Becker's (1981) term, she is an *effective altruist.* Only joint income matters for family decisions in these circumstances. Thus, effective altruism provides partial insurance for family members and insulates the family from targeted changes in taxes and benefits. Becker's (1981) claim that effective altruism also provides incentives for the beneficiary to act in the best interests of the family and reduces intra-family conflict – the so-called Rotten Kid Theorem – is, however, valid only with very restrictive preferences (Bergstrom 1989).

If, however, the couple's incomes are relatively similar, then neither spouse is rich enough relative to the other to make transfers to the other, and individual incomes are likely to affect family decisions. In either case, non-market interactions between family members are important for determining individual welfare, through either bargaining or intra-family transfers motivated by altruism, and these also affect market behaviour like consumer demand and labour supply.

Fertility, Investments in Children and Security in Old Age

The primary reason that most men and women enter a long-term relationship is to bear and raise children. In addition to the number of children, parents' welfare is likely to depend on the lifetime well-being of each child – 'child quality' for short. That is, parents receive more satisfaction from having children who are better off throughout their life, and they make monetary transfers and human capital investments to influence their children's lifetime standard of living.

If parents view child quantity and quality as substitutes and treat all their children equally, their budget constraint contains the product between the number of children and quality per child (Willis 1973). This implies that the 'shadow price' of an additional child is proportional to the level of child quality, and the shadow price of raising child quality is proportional to the number of children. As a consequence, there is an important interaction between family size and child quality. For example, a higher return to human capital increases investment per child. This raises the shadow price of children, which lowers family size and the price of child quality, thereby raising child quality further, and so on. Thus, increases in the returns to human capital investment associated with technical change may lead to simultaneous *large* reductions in fertility and increases in human capital investment in children. This is consistent with important stylized

facts of economic development and links family economics with the study of economic growth and development (Rosenzweig 1990).

The quantity–quality interaction may also produce a 'high fertility–low child investment trap', in which low quality produces a low price of children, high fertility and a high shadow price of child quality. Higher parents' income increases fertility and the price of child quality, keeping child investment low. It may take some policy or technological development that alters the prices of quantity or quality independently, such as a family planning intervention or a large change in the return to human capital investment, to 'spring the trap'. Once sprung, if the quality income elasticity exceeds the one for quantity, then the ratio of quality to the number of children rises with higher income, thereby increasing the shadow price of an additional child relative to the shadow price of child quality. The substitution effect induced by this increase may be sufficiently large to produce a decline in fertility when income increases, even though children are normal goods.

The ultimate manifestation of low child quality is a child not surviving to adulthood. Scientific advance or a policy intervention, such as better water supply or public health, increases the probability of child survival, but it has conflicting impacts on fertility. On the one hand, it reduces the price of a surviving birth, thereby encouraging higher fertility. But if parents can influence the chances that their own children survive to become adults by spending more on each child, then it is *possible* that exogenous improvements in child survival reduce fertility, provided that improvements in child survival substitute for parents' expenditure on child health (Cigno 1998). Such a relationship may help account for the 'demographic transition' – the change from a high fertility-high 'child mortality environment to a low fertility-low mortality one.

The factors affecting the cost of children (including investment in them) are closely associated with the key role of parental time in the rearing of and investment in children. The rearing of children is usually presumed to be *time-intensive* relative to other home production activities, and mothers provide a disproportionate share of parental time in the production of child quality. Thus, the cost of children relative to the cost of the parents' living standard is directly related to the mother's cost of time (Willis 1973). This links the cost of children with women's educational and earning opportunities, with implications for their effects on fertility, women's labour supply and investment in children.

Fertility and child investment may also be motivated by the need for support in old age. If people do not have access to a capital market, an extended family network including three generations at different stages of life could substitute for a capital market (Cigno 2000). In effect, it arranges 'loans' to its young members from its middle-aged ones and enforces repayment later when the young borrowers have become middle-aged and the middle-aged lenders have become old. People may have children only because they are needed to transfer resources through time. The opening of a capital market with a sufficiently high interest rate offers the middle-aged an alternative to this family transfer system. A threat of no support from the family in old age is no longer a deterrent, because they can make their own provision for old age through the market. In broad terms, this prediction is consistent with the observation that the growth of the financial sector, or the introduction of a state pension system, tends to coincide with a sharp decline in private transfers from the middle-aged to their elderly parents and a fall in fertility. The fact that childbearing does not cease suggests that the demand for children is not entirely derived from the need for transfers from them to finance consumption in old age. Again we see the important role of institutions in shaping family behaviour.

In countries with well-developed capital markets and pension systems, it is often observed that financial transfers from adult children to parents are rare, but transfers in the other direction are more common. Also, children are often observed providing 'services' to their parents that do not have clear market substitutes, such as companionship, attention and adapting their behaviour to their parents' wishes. Such services come at a cost to the children, and so transfers from parents to their child may be an exchange for these

services (Cox 1987). Parents may also want to help their children financially when they need it, but they want them to behave responsibly in the sense of expending sufficient effort to support themselves. How transfers from parents respond to an adult child's income depends on the balance of altruistic motives, parents' intention to provide an incentive for high effort and the effects of parent–child bargaining on the provision of child-services.

Marriage and Divorce

In addition to love and companionship, marriage offers two people the opportunity to share household public goods and benefit from the division of labour, and it facilitates risk sharing. Whom a person marries influences family behaviour (for example, fertility) and individual welfare through family resources, bargaining and costs (for example, of children). But the process of finding a spouse is one in which information is scarce, and it takes time to gather it. These market frictions affect who marries whom, the gains from each marriage and the distribution of gains between spouses (Burdett and Coles 1999). The positive correlation between spouses in desirable attributes like education is expected to be weaker when frictions are larger. The chances of divorce, and therefore divorce laws, also affect matching in the marriage market. A higher divorce rate makes people less choosy when selecting a spouse, because it reduces the perceived benefits from waiting for a better match by making it more likely that a person will return to the single state. Poorer matches ensue, and these have a higher probability of dissolving.

Marriage market frictions may also be responsible for childbearing outside marriage. A woman who has a relationship with a man she does not wish to marry, or who will not marry her, would choose to have a child by the man if the short-run gain exceeds the long-term costs in terms of damage to her marriage prospects (Ermisch 2003, ch. 7). Those women who expect to obtain a significant increase in welfare when they marry suffer a greater long-term cost by having a child while single than women whose marriage prospects are such that they expect to gain little from marriage. Thus, women with poorer marriage prospects are more likely to have children outside marriage.

Parents are likely to continue to care about the welfare of their children after they divorce, and so expenditure on children, such as investment in their human capital, is a public good to the parents. When living together, they choose the efficient level of this public good. But after breaking up, the mother usually obtains custody of the children and she decides the level of expenditure on children (Weiss and Willis 1985). The father can influence it only by making transfers to the mother, and he must transfer more than a dollar to obtain a dollar's more expenditure on children, because the mother spends part of the transfer on herself. The higher effective price for child expenditure when divorced encourages him to spend less on children after divorce (perhaps nothing), resulting in a lower, inefficient level of expenditure on children overall. This is likely to have implications for the lifetime welfare of children. The probability that a couple divorces is inversely related to this efficiency loss from divorce.

Behaviour within marriage is likely to be affected by exogenous variation in the probability of divorce (for example, through legal changes). If, for example, more participation in paid employment raises future wages, the risk of divorce can encourage more paid employment by the mother during marriage and, by raising the cost of child quality, lower expenditure on children and lower fertility. These 'defensive investments' are undertaken to increase welfare later, when outcomes are uncertain because of the possibility of divorce. Thus, the probability of divorce affects women's wages and labour supply in the economy.

Examples have illustrated the distinctive aspects of family economics: how market prices and personal incomes affect non-market interactions between individuals in the family (through altruistic motives and bargaining), fertility and investment in children. These channels link family economics to traditional fields like growth and development, labour economics, consumer demand, savings and inter-generational transfers.

See Also

- ▶ Child Health and Mortality
- ▶ Collective Models of the Household
- ▶ Family Decision Making
- ▶ Fertility in Developed Countries
- ▶ Fertility in Developing Countries
- ▶ Marriage and Divorce
- ▶ Rotten Kid Theorem

Bibliography

Becker, G. 1960. An economic analysis of fertility. In *Demographic and economic change in developed countries*, ed. National Bureau of Economic Research. Princeton: Princeton University Press.

Becker, G. 1981. *A treatise on the family*. Cambridge, MA: Harvard University Press.

Bergstrom, T. 1989. A fresh look at the Rotten Kid Theorem: And other household mysteries. *Journal of Political Economy* 97: 1138–1159.

Bergstrom, T. 1996. Economics in a family way. *Journal of Economic Literature* 34: 1903–1934.

Browning, M., and P.-A. Chiappori. 1998. Efficient intra-household allocations: A general characterization and empirical tests. *Econometrica* 66: 1241–1278.

Burdett, K., and M. Coles. 1999. Long-term partnership formation: Marriage and employment. *Economic Journal* 109: F307–F334.

Cigno, A. 1998. Fertility decisions when infant survival is endogenous. *Journal of Population Economics* 11: 21–28.

Cigno, A. 2000. Self-enforcing family constitutions. In *Sharing the wealth: Intergenerational economic relations and demographic change*, ed. A. Mason and G. Tapinos. Oxford: Oxford University Press.

Cox, D. 1987. Motives for private income transfers. *Journal of Political Economy* 95: 508–546.

Ermisch, J. 2003. *An economic analysis of the family*. Princeton: Princeton University Press.

Konrad, K., and K. Lommerud. 1995. Family policy with non-cooperative families. *Scandinavian Journal of Economics* 97: 581–601.

Pollak, R. 1985. A transaction cost approach to families and households. *Journal of Economic Literature* 23: 581–608.

Rosenzweig, M. 1990. Population growth and human capital investments: Theory and evidence. *Journal of Political Economy* 98: S38–S70.

Weiss, Y., and R. Willis. 1985. Children as collective goods and divorce settlements. *Journal of Labor Economics* 3: 268–292.

Willis, R. 1973. A new approach to the economic theory of fertility behavior. *Journal of Political Economy* 81: S14–S64.

Family Planning

Mark R. Rosenzweig

The phrase ‘family planning’ has come to mean the set of institutions, policies and programmes whose principal objective is to alter the family size decisions of households. Family planning institutions, private or public, attempt to influence fertility choices by (a) direct persuasion of couples to adopt socially ‘appropriate’ family size goals; (b) the dissemination of information on techniques of birth or conception prevention, and (c) the provision of birth or conception control services or inputs at subsidized cost. In addition, governments may adopt policies that directly alter the incentives for bearing and rearing children. Such policies may include income tax exemptions or direct transfers which vary by the number of children and/or economic and social sanctions related to family size, such as restrictions on parental work opportunities or restrictions on schooling or consumption privileges when those are principally supplied by the public sector.

Of course, to the extent that fertility decisions are responsive to changes in relative prices and to income changes, all governmental policies (tax, transfer, expenditures) indirectly influence the family size goals of households. What principally distinguishes family planning interventions from other government programmes is their attempt to affect fertility outcomes by influencing the means by which households achieve their family size goals.

Family Planning and the Economic Theory of Fertility

Economic models of fertility that incorporate the technology of reproduction provide a general framework with which to analyse the influence of family planning programmes on the family size plans of families (Easterlin et al. 1980; Rosenzweig and Schultz 1985). In these models,

births or conceptions are viewed as byproducts of sexual activity. These byproducts can be averted by the employment of methods of birth control or contraceptive techniques. The set of relationships between sexual and other behaviours, contraceptive practices, and conceptions or births is the reproductive technology, analogous to the technology of production in firms, which describes the effects of inputs on outputs. Couples thus determine their fertility through the use of reproduction inputs. And just as firms adjust output when either input prices or the technology of production change, given demand for the firm's product, couples alter their fertility in response to changes in the costs of reproductive inputs or to changes in the technology of reproduction, given their family size goals.

Family planning initiatives that lower the costs of averting births through subvention of reproduction inputs or information provision have price and income effects. The lowering of the costs of averting births induces couples to avert more births (the own price effect); but couples' real incomes are also higher as a consequence and they may decide to spend some of that income by having larger families. If income effects are small relative to price effects (more likely the smaller the share of contraceptive costs in the family budget), such family planning activities should lower fertility, whatever the motivations of couples for having children.

The degree to which a couple benefits from or is influenced by programmatic family planning activities depends on its family size goals and on the type of family planning activity. If family planning interventions make birth reduction less costly, those couples who desire smaller families (avert more births) benefit most. If the poorest households have the largest families it is thus not clear that non-selective contraceptive *subsidy* programmes benefit the poor relative to the rich. To the extent, however, that family planning initiatives are characterized chiefly by information dissemination, the distribution of the benefits will depend on the pre-programme distribution of such information in the population. If more educated or wealthier couples are better able to acquire information in the absence of such programmes than are other couples, the programmes will benefit such couples least. Fertility reductions associated with contraceptive information dissemination will be larger in poorer, less-educated families.

Economic theory also suggests that the effects of family planning, by altering the costs of fertility, will not be confined to changes in family size. As noted, the increase in income associated with the subsidy may be spread among other family activities. But there are also substitution or cross price effects. In models (Willis 1973; Becker and Lewis 1973) in which couples care about the average 'quality' of children, reductions in the cost of fertility control and thus reduced family size make the provision of resources to children less costly, as such resources need be allocated among less children. If family size and child quality are substitutes in the usual consumer demand sense, then it is likely that the reduction in fertility induced by family planning activities will also result in increased investments by families in each child born even if there are no direct biological links between birth order, birth intervals and the characteristics of children.

Rationales for Family Planning Interventions

Rigorous theoretical justifications for the public subvention of family planning activities are surprisingly scarce. As for all public interventions, a rationale on efficiency grounds should be based on a demonstration that the costs incurred by private agents making fertility decisions diverge from the social costs of those decisions. The exact nature of the market failure or market incompleteness or the direct negative externalities associated with the production of children that might render family planning programmes appropriate instruments for achieving more efficient outcomes in an economy have not been clearly identified. In growth models incorporating optimal fertility decision-making, the results appear to depend critically on the assumed degree of altruism parents have for children (and vice versa), the allocation of property rights over parental investments

in children, and the completeness of intertemporal markets. In the absence of clear resolutions of such issues, a number of other justifications for publicly supported family planning activities have been put forth. One rationale is based on the existence of positive externalities associated with human capital investment (Rosenzweig and Wolpin 1986). If investments in health or in schooling by households directly benefit other households such that public subventions of such activities are optimal, then it may be efficient to subsidize fertility control (a) if reductions in family size induce greater investments in human capital and/or (b) since reductions in the number of children make less costly public subsidization of investments in children. This argument suggests that health, schooling and family planning programmes are complementary and would tend to be positively correlated over time within countries and across areas.

Two other rationales for family planning interventions are based on information problems. The rise in incomes accompanying economic development and the use of newer medical technologies have contributed to the dramatic fall in infant and child death rates in low income countries over the past decades without a concomitant decline in fertility in many countries. If parents do not correctly foresee the future drop in the risk of death for their children associated with the health externalities of economic growth and development (infection reduction), then subsidization of fertility control may be warranted to reduce fertility to appropriate levels.

Technological innovation has also characterized the control of fertility. If the market provision of information about new methods of contraception is problematic, then publicly funded information dissemination about innovations in this technology may be warranted. Family planning services are then analogous to extension services in agriculture.

Evaluating Family Planning Programmes

The conceptual experiment needed to ascertain how and to what extent family planning subsidies or information provision actually influence fertility and other behaviours is straightforward - randomly select an area or set of areas for intervention and compare the fertility and other relevant outcomes there with those in non-intervention areas. Since dynamic models of fertility (e.g. Heckman and Willis 1978) have as yet little to say about how reductions in the costs of fertility control influence the timing and spacing of births, it may not be appropriate to measure the effects of such programmes over short intervals of time. Couples with less costly and/or improved means of controlling fertility may choose to have their children earlier or later; the short-run response of fertility to a family planning intervention may be quite different from the response in terms of completed family size.

Information from appropriate randomized experiments involving family planning activities is scarce. Most estimates of the impact of family planning interventions have come from non-experimental data, chiefly cross-sectional data. The best of the cross-sectional studies of the effects of public expenditures on family planning or measures of access to family planning institutions examine as well the natalist effects of other programmes (health programmes, for example). Since theory suggests that health and family planning interventions are complementary and are likely to be distributed similarly, failure to take into account the existence and distribution of other programmes when evaluating the impact of family planning interventions may yield misleading estimates of family planning efforts. Multivariate studies combining spatial information on programmes and household data from rural and urban Colombia and rural India (Rosenzweig and Schultz 1982; Rosenzweig and Wolpin 1982) indicate that family planning and health institutions (clinics) are associated with both lower fertility and lower rates of child mortality, although no effects of these programmes were found in rural areas of Colombia. Results from the urban Colombia data, moreover, indicated that the effects of the programmes were significantly greater among households with less-educated mothers. This result is consistent with the notion that the family planning (and health) programmes

principally serve to disseminate information, this function being of less value for the more educated (and better informed) households.

A study using longitudinal information on the nutritional status of children and information on the dates of initiation of health and family planning programmes (Rosenzweig and Wolpin 1986) tested whether the timing of public programme interventions across areas was correlated with unmeasured area factors associated with child health. The results suggested the spatial distribution of both health and family planning programmes was not random, with both types of programmes tending to be similarly placed (in low health areas), and that once non-random programme placement was taken into account (but not before), both the family planning and health programmes appeared to improve significantly the nutritional status of children.

These empirical studies thus suggest that family planning activities have succeeded in lowering fertility and in augmenting human capital investment, in at least some countries, but that more attention to the rules by which public programmes are distributed and initiated may be needed to obtain more accurate estimates of the effects of such programmes. Improved estimates of the consequences of family planning initiatives are thus a byproduct of a better understanding of the rationale for such programmes and of public sector behaviour.

See Also

- ► Fecundity
- ► Fertility

Bibliography

Becker, G.S., and H.G. Lewis. 1973. On the interaction between quantity and quality of children. *Journal of Political Economy* 82(April/May): S279–S288.

Easterlin, R.A., R.A. Pollak, and M.L. Wachter. 1980. Toward a more general economic model of fertility determination. In *Population and economic change in developing countries*, ed. R.A. Easterlin. Chicago: University of Chicago Press.

Heckman, J.J., and R.J. Willis. 1978. Estimation of a stochastic model of reproduction: An econometric approach. In *Household production and consumption*, ed. N.E. Terlecky. New York: Columbia University Press.

Rosenzweig, M.R., and T.P. Schultz. 1982. Child mortality and fertility in Colombia: Individual and community effects. *Health Policy and Education*, February, 125–151.

Rosenzweig, M.R., and T.P. Schultz. 1985. The demand for and supply of births: Fertility and its life-cycle consequences. *American Economic Review* 75(December): 992–1015.

Rosenzweig, M.R., and K.I. Wolpin. 1982. Governmental interventions and household behavior in a developing country: Anticipating the unanticipated consequences of social programs. *Journal of Development Economics* 10(2): 209–225.

Rosenzweig, M.R., and K.I. Wolpin. 1986. Evaluating the effects of optimally distributed public programs: Child health and family planning interventions. *American Economic Review* 76(June): 470–482.

Willis, R.J. 1973. A new approach to the economic theory of fertility behavior. *Journal of Political Economy* 81(March/April): S14–S64.

Famines

Cormac Ó Gráda

Abstract

Today the ultimate Malthusian check of 'gigantic, inevitable famine' is confined to the very poorest pockets of the globe. Economic development, medical technology and the globalization of disaster relief have reduced the size and duration of famines in the recent past. On the other hand, totalitarianism and the enhanced role of human agency produced in the 20th century some of the biggest famines ever. Topics discussed include the demography and long-run impact of famine, the role of public and private action in relieving those at risk, and how markets function during famines.

Keywords

Agency problems; Charitable donations; Corruption; Democracy; Demographic transition; Epidemiological transition; Famines; Fertility;

Food availability declines; Food markets; Global warming; Health; Hoarding; Life expectancy; Malthus, T.R.; Migration; Mortality; Non-governmental organizations; Nutrition; Poverty; Public works; Sen, A.K

JEL Classifications
J10

'Famine' is defined narrowly here as a food shortage leading directly to excess mortality from starvation or hunger-induced illnesses (compare Howe and Devereux 2004). By this definition, the 20th century presents a paradox in the history of famines. On the one hand, it witnessed in China in 1959–61 the greatest famine in world history. On the other, it saw the virtual elimination of famine across most of the globe. Economic growth in the 19th century led to the disappearance of famine in Europe in peacetime and, after the 1870s, a reduction in famine intensity throughout Asia.

Today's high-profile famines are, relatively speaking, small and confined to poverty-stricken and often war-torn corners of Africa. In principle, famine prevention should be 'easy'. Better communications, better understanding of nutritional requirements and medical remedies, and the globalization of disaster relief mean that the risks faced by the world's most underdeveloped economies should be far fewer than those faced by equally poor countries in the past.

'Malthusian' Famines

Famines and economic backwardness are closely related. Malthus would not have been surprised to hear of famine in Niger, probably the world's poorest economy, in 2005, or that the cross-sectional correlation between excess mortality and poverty was strong within Ireland in the 1840s and Bengal in the 1940s. And he would have deemed the extreme backwardness of the Chinese economy in the mid-1950s a contributory factor to the Great Leap Forward famine of 1959–61: Chinese real GDP per head then was less than half the African average in 2006 (Maddison 2006).

Most famine victims succumb to infectious diseases rather than to actual starvation. Poverty prevents proper medical care because the associated remedies are costly and difficult to implement in crisis conditions. Sub-Saharan Africa has yet to complete the 'epidemiological transition', mainly because the resources and the political capabilities to put what is available locally or obtainable from abroad to most effective use are lacking. Famines are the exception where the transition has been completed, but when they occur, as in Nazi-occupied Leningrad, Greece, and the western Netherlands during the Second World War, the diseases mainly responsible for excess mortality were very different. In these relatively developed societies, public health structures that prevented the spread of infectious disease had become part of daily routine, and continued to be so during the war (Mokyr and Ó Gráda 2002; Maharatna 1996, pp. 159–61; Hionidou 2006).

Most famines strike in the wake of major crop failures, although crop failure is neither a necessary nor a sufficient condition for famine. Even the most backward economies often have the resilience to cope with once-off harvest shortfalls, so that in the past the worst famines have been the product of back-to-back shortfalls of the staple crop. Thus, the probability of back-to-back poor harvests should provide some sense of the likelihood of famine in the past. Agricultural and meteorological data imply that such back-to-back events were uncommon (Ó Gráda 2007).

Entitlements and Governance

Civil unrest and bad government can also lead to famine by limiting production and trade or failing to prevent the spread of epidemic disease. The impact of war on the supply of shipping and grain imports from abroad was an important contributory factor to famine in Bengal in 1943–44. Panics about the food supply and poorly

performing food markets may exacerbate famine. In such instances factors other than crop shortfalls reduce the purchasing power or 'entitlements' of vulnerable sections of the population: the size of the loaf matters less than its distribution. Claims that even during famines there is adequate food for everyone are not new. Such claims, which invert the relative importance of food supply on the one hand, and human action and distribution on the other, had a particular resonance for the 20th century.

On several occasions between the 1930s and the 1950s, not only did totalitarian regimes engage in policies that placed millions at risk, but they also managed to keep the consequences largely hidden from the outside world. Analyses of 20th-century famines accordingly have tended to dwell less on economic factors such as the background level of development and the extent of the crop shortfall than on the role of human agency – be it the ruthlessness of dictators or the incompetence of officialdom. Yet closer inspection suggests that even the most notorious 'man-made' famines of the 20th century in the Soviet Union in 1932–33, in China in 1959–61, and even in Bengal in 1943–44, entailed what Amartya Sen (1981) has dubbed 'food availability declines' (FADs) (Davies and Wheatcroft 2004: ch. 5; Tauger 2006; Ó Gráda 2007). The paucity of evidence for 'pure' entitlement famines – famines where there was no food availability decline – suggests that modern scholarship may underestimate the role of food supply in the relatively recent past.

Sen's claim that famine and democracy are incompatible (Sen 2001) is a special case of the more general claim that democratic institutions promote economic justice and reduce inequality. Exceptions to this rule seem few: Banik's analysis of press reports of starvation deaths in Orissa in the 1990s confirms it in so far as famines are concerned, but highlights the inability of a free press and collective action to prevent mass malnutrition and 'many, many deaths' (Banik 2002). It also bears noting that in poverty-stricken, ethnically divided, low-literacy economies democracy may not be sustainable. Nonetheless, the exogenous element in democratic institutions surely matters.

Markets and Famines

Economists have long argued that, since crop failures are subject to spatial variation and rarely occur two years in succession, spatial and intertemporal arbitrage in food markets should help mitigate the cost of famines (Persson 1999). However, natural obstacles (poor communications) and artificial obstacles (war, civil unrest, trade restrictions and price controls) have often impeded the scope for arbitrage.

Research on Bengal in 1942–44 and Bangladesh in 1974–75 claims that food markets worked poorly in these instances, in the double sense of inadequate interregional trade and 'excessive' hoarding on the part of producers and traders (Sen 1981; Ravallion 1987, pp. 19, 111–13; 1997, pp. 1219–21). Formal studies of market performance during pre-20th century famines are few, although evidence from pre-industrial Europe suggests that they functioned no worse than in normal times (Ó Gráda 2005). The asymmetry in speculators' expectations implied by the findings of Sen and Ravallion – over-pessimism in the event of a harvest shortfall – is absent in the earlier data. That does not mean that markets worked like clockwork in pre-industrial Europe, but merely that their responses to spatial and intertemporal disequilibria were no weaker than in non-crisis times. In practice, markets may adjust too slowly to prevent famine: in the mid-19th century, for example, before the telegraph and long-distance bulk carriage by steamship could have made the difference, global grain markets could not have prevented mass mortality in Ireland and India. Nor does this mean that well-functioning, integrated markets always benefit the poor: as Sen emphasizes, they might allow inhabitants of less affected areas, endowed with the requisite purchasing power, to attract food away from famine-threatened areas. Much depends on whether such exports are used to finance cheaper imported substitutes, and on the speed with which food markets adjust. Dogmatic generalizations are not warranted.

Free markets can mitigate the impact of famines in two other respects. First, migration arguably limits the damage wrought by poor harvests,

since the migrants reduce the pressure on scarce food and medical resources where the crisis is deepest. This is probably true even when the poorest lack the resources to migrate. Although migration undoubtedly exacts a cost in terms of the spread of infectious disease in host countries, on balance it saves lives.

Second, regional specialization increases aggregate output, with a resultant reduction in the risks attendant on any proportionate harvest shortfall. Increasing commercialization also makes for more effective arbitrage in food markets. For example, the implied reduction in the cost of holding carry-over stocks and of transport greatly reduced the vulnerability of the Italian and the English poor in the early modern era (Persson 1999; Ó Gráda 2007).

Public and Private Action

Throughout history, whether out of fear or compassion, ruling elites have accepted a degree of responsibility for those at risk during famines. Most analytical attention has focused on the management rather than the extent of relief allocation. Since human interventions almost always give rise to principal–agent problems, choosing the appropriate yardstick for effective famine relief is an abiding issue. In the past, because governing elites were remote from those at risk, they often relied on sub-bureaucracies and landowners to identify deserving recipients of relief. History is full of examples of trade-offs between the degree of delegation on the one hand, and corruption and red tape on the other (see, for example, Shiue 2004).

The choice of appropriate public action in the presence of such agency problems during famines is discussed in Drèze and Sen (1989), Besley and Coate (1992), Ravallion (1997), and elsewhere. Transfers of food at subsidized prices may risk corruption and hoarding; hence the frequent focus on the provision of nontradable and highly perishable food rations. Income transfers (for example, through wages paid on public work schemes) are less likely to distort food markets, though if linked to work performance they may well discriminate against those in most need. Public works schemes also risk spreading infectious diseases. A further problem with public works is that fiscal stringency or fears of distorting labour markets, as in Ireland in the 1840s and in southern India in the 1870s, may entail below-subsistence wages and consequent excess mortality.

Private charity can mitigate famine but is rarely adequate during big crises.

Since the 1950s famine relief has been globalized through non-governmental organizations (NGOs) such as Oxfam. NGOs have been effective at highlighting the link between Third World poverty and the risk of famine, and at fund-raising in the wake of highly publicized crises. Nonetheless, their record in mitigating and averting famine raises several issues.

First, agencies originally founded as famine relief agencies tend to reinvent themselves as bureaucracies. Such organizations must balance the public's wish to relieve disasters as they happen with their own need for bureaucratic sustainability. This has entailed focusing more on development than on famine relief per se. Budgetary pressures have also tempted NGOs to exaggerate the risks or gravity of famine, or to claim the credit when the crisis is 'averted' (De Waal 1997). Given the likely long-term costs of such tactics, and the recent increasing dependence of NGOs on public funding, independent monitoring of their activities is essential. Moreover, NGO interventions typically lag, rather than lead, media reports; instead of drawing on previously accumulated reserves, they rely on crises to solicit aid, and their overreliance on emergency-generated funding has led them to locations where they lack the detailed expertise and connections essential for effective famine relief. Most NGOs continue to spread themselves too thin, and are too small to offer the insurance required for a rapid response against famine.

Measuring the Demographic Cost

Soaring food prices and poor harvests are often harbingers of famine, but are neither necessary nor sufficient conditions for one. On the one

hand, appropriate relief policies may prevent famine; on the other, not all famines result from aggregate food deficits or inflated food prices. An abnormal jump in mortality is a surer signal of famine, and is usually regarded as its defining feature. For most historical famines, however, establishing excess mortality with any precision is impossible, and inferences derived from incomplete data are often controversial. Much hinges on assumptions about the under-registration of deaths at the time. Controversy still surrounds the true tolls in the Soviet Union in 1931–33, China in 1959-61, Cambodia in 1975–79, and North Korea in 1995–99.

Nonetheless, it is clear that modern famines are, relatively speaking, far less costly in terms of human lives than earlier famines. Although non-crisis death rates in Africa remain high, excess mortality from famines in recent decades has been low. In Devereux's useful listing of major 20th-century famines only two – Nigeria in 1968–70 and Ethiopia in 1983–85 – are accorded tolls nearing one million (Devereux 2000). Elsewhere, deaths were far fewer.

Although famine had virtually disappeared from Europe by the mid-19th century, 30 million is a conservative estimate of famine mortality in India and China alone between 1870 and about 1900, and 'fifty million might not be unrealistic' (Davis 2001, p. 7). One hundred million would be a conservative guess at global famine mortality during the 19th century as a whole. Given that global population rose from about 1.3 billion in 1870 to 2.5 billion in 1950, in relative terms famines were much more lethal in the 19th century than in the 20th. The late 19th century saw a reduction in famine intensity in India, due to a combination of better communications and improvements in relief policy; in Russia, too, famines became more localized. Japan, where famines were common in the 17th century, and less so in the 18th, experienced its last true famine in the 1830s.

As noted earlier, infectious diseases usually account for most famine deaths. These include deaths due to diet-related diseases brought on by impaired immunity, or to poisoning from inferior or unfamiliar foods. They also include deaths stemming from the disruption of personal life and societal breakdown attendant on famine. Disease spreads with the increased mobility of the poor and the inevitable deterioration in sanitary conditions. Famines also are associated with outbreaks of seemingly unrelated diseases such as cholera, influenza, and malaria (Mokyr and Ó Gráda 2002).

The implications of focusing on relative rather than absolute mortality are also worth noting. In relative terms, excess mortality in China in 1959–61 was modest compared, for example, with Ireland in the 1840s or Finland in 1867–68. The lower rate matters to the extent that it affected the characteristics of the famine. But such comparisons beg the question of the appropriate denominator. Most of these famines were regionally concentrated, but the denominators refer to larger political or geographical units. Finally, most famines last a year or two at most. Ireland in the 1840s, Cambodia in the 1970s, and North Korea in the 1990s are exceptional in this respect.

Although in the past non-crisis male life expectancy usually exceeded female, the evidence for a female advantage during famines is overwhelming (for example, Hionidou 2006, p. 165; Maharatna 1996, pp. 231–4). The main reason for this is physiological. Whether the female advantage has changed over time remains a moot point, but there is some presumption that the female advantage is greater when the main cause of death is literal starvation. Most famine victims tend to be the very young and those beyond middle age, although the greatest *proportional* increases in death rates are at ages in between. In cases where population growth of two or three per cent per annum is the norm, such age and gender biases are unlikely to have much impact, and population growth may be expected to quickly fill the resultant demographic vacuum. Where non-crisis growth is slow, these biases may matter more, and post-famine recovery is likely to be slower.

For several reasons, the demographic consequences of famine are more complex than implied by the standard measure of excess mortality. First, that measure ignores the drop in births that usually accompanies famine. Famines almost

invariably entail significant reductions in births and marriages (for example, Maharatna 1996, pp. 179–83; Hionidou 2006, pp. 178–89). There is a case for including the births deficit in the demographic reckoning. Births lost due to the Great Irish Famine numbered about 0.4 million in a population of eight million, whereas estimates for China in the wake of the 1959–61 famine run as high as 30 million in a population of 650 million (Yao 1999). There are several reasons for such declines in the birth rate, including lower libido, spousal separation, and weaker reproductive functioning. Famines also usually entail fewer marriages although, clearly, in most situations marriage reductions have implications only for *first* births.

Second, the excess mortality measure omits both the rebound in the birth rate and the decline in the death rate that sometimes follow once the crisis has passed. Births in China in 1962 exceeded those in any year since 1951, and in the following three years the birth rate was also higher than in any other year in the 1950s and 1960s. Therefore, to some extent at least, births 'lost' during the famine seem to have been merely postponed.

Third, it leaves out of account any longer-run impact on mortality and morbidity. Famines hasten the deaths of some ill and elderly people who would have died soon in any case. The ensuing impact on the demographic structure entails a reduction in the death rate in the wake of famines.

Long-Term Health Effects

Recent medical-historical research has revealed a close link between health and nutrition *in utero* and in early childhood on the one hand, and adult health and longevity on the other (Barker 1992). The implications for the long-term demographic and health effects of famines are obvious. Research on Russian, Dutch and Chinese data links foetal exposure to famine to increased risks in later life of diseases as varied as schizophrenia, breast cancer, arteriosclerosis, and antisocial personality disorders (Khoroshinina 2005, p. 208). There is evidence from Leningrad that being born just before or during famines reduces expected adult height (for example, Kozlov and Samsonova 2005, pp. 178–89; Khoroshinina 2005, pp. 198–200). Such evidence suggests that the human cost of famines has been underestimated in the past, although it is too soon to say by how much. Finally, there is the further disturbing possibility – still unexplored – that famine-induced malnutrition *in utero* or early childhood adversely affects the mental development of those at risk.

Conclusion

Famine's range has been narrowing since Malthus's time. By 1900 Europe and its industrialized extensions, Latin America, and Japan were virtually famine-free, and today major, *prolonged* famine anywhere is conceivable only in contexts of endemic warfare or self-enforced isolation. Compared with the persistent effects of HIV/AIDS on the population of sub-Saharan Africa, the damage wrought by famine is minimal. Moreover, given that throughout most of history land hunger has been a powerful predictor of famine, recent trends in the balance between population and food production offer room for cautious optimism about the near future. In both Asia and Latin America, food production has grown much faster than population since the 1960s. In sub-Saharan Africa the balance has been much closer, although the problem there has been very rapid population growth rather than sluggish food output growth. Moreover, some African countries such as Burkina Faso and Niger have walked a high demographic tightrope while others (such as Malawi and Zimbabwe) have performed poorly despite slower population growth.

The few remaining places still vulnerable to textbook Malthusian famine are those yet to undergo the fertility decline of the demographic transition. Those countries have experienced considerable mortality improvement in recent decades, but they lag behind in terms of fertility decline. A key issue is how fertility decline, scarcely yet under way, unfolds in such vulnerable economies. The experience of post-fertility transition economies worldwide strongly mirrors the

historical pattern whereby declines in fertility were preceded by declines in mortality. However, the length of the lag and the extent of the fertility decline are clearly crucial. A guarded historical lesson for countries like Niger is that the transition, once under way, has been more rapid in latecomers than in pioneers. Africa's sluggish fertility transition, itself a function of economic underdevelopment, has increased its share of global population from only 8.8 per cent in 1950 to 14 per cent today; it is set to reach 21.7 per cent by 2050. Even though a drop in the annual growth rate from 2.5 per cent during the second half of the 20th century to 1.4 per cent during the first half of the 21st in Africa as a whole is implied, population is predicted to treble by 2050 in famine-prone countries such as Niger, Uganda and Mali. When coupled with the problem of global warming, which is likely to impact disproportionately on the productivity of arid lands limited to a short growing season, the implied threat to living standards is clear.

See Also

- ▶ Malthus, Thomas Robert (1766–1834)
- ▶ Poverty
- ▶ Sen, Amartya (Born 1933)

Bibliography

Banik, D. 2002. Democracy, drought and starvation in India: Testing Sen in theory and practice. PhD thesis, Department of Political Science, University of Oslo.

Barber, J., and A. Dzeniskevich, eds. 2005. *Life and death in Leningrad, 1941–44*. London: Palgrave Macmillan.

Barker, D.J.P., ed. 1992. *Fetal and infant origins of adult disease*. London: BMJ Publishing Group.

Besley, T., and S. Coate. 1992. Workfare vs. welfare: Incentive arguments for work requirements in poverty alleviation programs. *American Economic Review* 82: 249–261.

Davies, R.W., and S.G. Wheatcroft. 2004. *The years of hunger: Soviet agriculture, 1931–33*. London: Palgrave Macmillan.

Davis, M. 2001. *Late victorian holocausts*. London: Pluto.

Devereux, S. 2000. Famine in the twentieth century. Working Paper No. 105, Institute of Development Studies, University of Sussex.

De Waal, A. 1997. *Famine crimes: Politics and the disaster relief industry in Africa*. *Oxford*: *James Currey.*

Drèze, J., and A. Sen. 1989. *Hunger and public action*. Oxford: Oxford University Press.

Dyson, T., and C. Ó Gráda, eds. 2002a. *Famine demography: Perspectives from the past and present*. Oxford: Oxford University Press.

Hionidou, V. 2006. *Famine and death in occupied Greece, 1941–1944*. Cambridge: Cambridge University Press.

Howe, P., and S. Devereux. 2004. Famine intensity and magnitude scales: A proposal for an instrumental definition of famine. *Disasters* 28: 353–372.

Khoroshinina, L. 2005. Long-term effects of lengthy starvation in childhood among survivors of the siege. *Barber and Dzeniskevich* 2005.

Kozlov, I., and A. Samsonova. 2005. The impact of the siege on the physical development of children. *Barber and Dzeniskevich* 2005.

Maddison, A. 2006. World population, GDP and per capita GDP, 1–2003 AD (2006 update). Online. Available at http://www.ggdc.net/maddison/. Accessed 31 Jan 2007.

Maharatna, A. *1996*. *The demography of famines: An Indian historical perspective*. *Delhi*: *Oxford University Press*.

Mokyr, J., and C. Ó Gráda. 2002. What do people die of during famines? The great Irish famine in comparative perspective. *European Review of Economic History* 6: 339–364.

Ó Gráda, C. 2005. Markets and famines in pre-industrial Europe. *Journal of Interdisciplinary History* 26: 143–166.

Ó Gráda, C. 2007. Making famine history. *Journal of Economic Literature* 31: 3–36.

Persson, K.-G. 1999. *Grain markets in Europe 1500-1900, integration and regulation*. Cambridge: Cambridge University Press.

Ravallion, M. 1987. *Markets and famines*. Oxford: Oxford University Press.

Ravallion, M. 1997. Famines and economics. *Journal of Economic Literature* 35: 1205–1242.

Sen, A. 1981. *Poverty and famines: An essay on entitlement and deprivation*. *Oxford*: *Oxford University Press*.

Sen, A. 2001. *Development as freedom*. Oxford: Oxford University Press.

Shiue, C.H. 2004. Local granaries and central government disaster relief: moral hazard and intergovernmental finance in eighteenth- and nineteenth-century China. *Journal of Economic History* 64: 100–124.

Tauger, M. 2006. Arguing from errors: on certain issues in Robert Davies' and Stephen Wheatcroft's analysis of the Soviet grain harvest and the Great Soviet Famine of 1931–33. *Europe–Asia Studies* 58: 973–984.

Yao, S. 1999. A note on the causal factors of China's famine in 1959–1961. *Journal of Political Economy* 107: 1365–1369.

Fannie Mae, Freddie Mac and the Crisis in US Mortgage Finance

Lawrence J. White

Abstract

Fannie Mae and Freddie Mac are two large companies – 'government-sponsored enterprises' (GSEs) – that are heavily involved in the secondary market for residential mortgages. The GSEs' expansion into lower quality mortgages, especially during the middle years of the 2000s, was supported by insufficient capital and led to their insolvency and conservatorships on 6 September 2008 – which essentially placed them under full government control. As of the spring of 2011 they remain as mainstays of the US residential mortgage market; but they also remain in conservatorships. Their future and the future of mortgage finance is an active topic of political debate.

Keywords

Capital; Fannie Mae; Freddie Mac; Government-sponsored enterprises; Implicit guarantee; Leverage; Mortgage-backed securities; Mortgage finance; Secondary mortgage market

JEL Classifications

G18; G21; G28; L85

Introduction

The Federal National Mortgage Association (more commonly known as 'Fannie Mae') and the Federal Home Loan Mortgage Corporation ('Freddie Mac') are two large companies – frequently described as 'government-sponsored enterprises' (GSEs) – that are heavily involved in the secondary market for residential mortgages. They played a major role in the expansion of residential mortgage finance in the 1990s and into the middle of the decade of the 2000s. When housing prices began to fall after mid-2006, mortgage borrowers began to default. The GSEs' expansion into lower quality mortgages, especially during the middle years of the decade of the 2000s, was supported by insufficient capital and led to their insolvency in the late summer of 2008 and to the US Government's decision to place them into conservatorships on 6 September 2008 – which essentially placed them under full government control.

As of the spring of 2011 they remain as mainstays of the US residential mortgage market; but they also remain in conservatorships. The Obama administration in February 2011 proposed a number of alternative structures for the future of residential mortgage finance, all of which involve the eventual demise of the two companies; but Congress has yet to take any action.

What They Do

Fannie and Freddie's business activities can be separated into two somewhat related functions:

1. They invest in residential mortgages. In essence, they buy mortgages from originators (i.e. from the entities that, in the first instance, lend to the mortgage borrower) and hold those mortgages on their own balance sheets. As of year-end 2009, Fannie Mae had $745 billion in mortgage assets on its balance sheet; Freddie Mac had $717 billion (see Table 1). Even before their insolvencies, they financed their holdings of mortgages almost entirely with debt; typically, $100 in mortgages would be financed with $96–$97 of debt and only $3–$4 of equity capital. They were thus highly leveraged.
2. They securitize residential mortgages. In this function, they buy mortgages from originators, bundle them into packages or 'pools' of mortgage-backed securities (MBS), and sell the MBS to investors (banks, insurance companies, pension funds, mutual funds, hedge

Fannie Mae, Freddie Mac and the Crisis in US Mortgage Finance, Table 1 Mortgages held and MBS outstanding, by Fannie Mae and Freddie Mac, 1948–2009 (all dollar amounts are in $ billions). All mortgage amounts encompass single-family mortgages plus multi-family mortgages (Sources: Federal Reserve 'Flow of Funds', various years; FHFA (2010))

Fannie Mae			Freddie Mac			
Year	Mortgages held in portfolio ($)	MBS outstanding ($)	Mortgages held in portfolio ($)	MBS outstanding ($)	Total US residential mortgages ($)	Total (F + F)/total residential mortgages (%)
1948	0.2				39.8	0.5
1949	0.8				45.2	1.8
1950	1.3				54.3	2.4
1951	1.8				62.3	2.9
1952	2.2				69.9	3.1
1953	2.5				78.1	3.2
1954	2.4				88.0	2.7
1955	2.6				101.4	2.6
1956	3.1				112.8	2.7
1957	4.0				121.9	3.3
1958	3.9				133.7	2.9
1959	5.3				148.7	3.6
1960	6.2				162.1	3.8
1961	6.1				177.6	3.4
1962	5.9				195.0	3.0
1963	4.7				215.1	2.2
1964	4.4				136.9	3.2
1965	4.7				257.6	1.8
1966	7.1				274.0	2.6
1967	8.9				290.7	3.1
1968	7.1				311.1	2.3
1969	11.0				331.8	3.3
1970	15.5				352.2	4.4
1971	17.9		0.9	0.1	388.5	4.9
1972	19.7		1.7	0.4	440.2	5.0
1973	23.6		2.5	0.8	493.0	5.5
1974	28.7		4.5	0.8	535.1	6.4
1975	30.8		4.9	1.6	574.6	6.5
1976	31.8		4.2	2.8	640.9	6.1
1977	33.3		3.2	6.8	742.0	5.8
1978	42.1		3.0	12.0	863.4	6.6
1979	49.8		4.0	15.3	990.7	7.0
1980	55.6		5.0	17.0	1100.4	7.1
1981	59.6	0.7	5.2	19.9	1172.6	7.3
1982	69.4	14.5	4.7	43.0	1216.3	10.8
1983	75.2	25.1	7.5	57.7	1347.3	12.3
1984	84.1	35.7	10.0	70.0	1507.2	13.3
1985	94.6	54.6	13.5	99.9	1732.1	15.2
1986	94.1	95.6	13.1	169.2	2068.8	18.0
1987	93.7	135.7	12.4	212.6	2186.1	20.8
1988	100.1	170.1	16.9	226.4	2436.6	21.1
1989	108.0	216.5	21.4	272.9	2655.9	23.3

(*continued*)

Fannie Mae, Freddie Mac and the Crisis in US Mortgage Finance, Table 1 (continued)

Fannie Mae			Freddie Mac			
Year	Mortgages held in portfolio ($)	MBS outstanding ($)	Mortgages held in portfolio ($)	MBS outstanding ($)	Total US residential mortgages ($)	Total (F + F)/total residential mortgages (%)
1990	114.1	288.1	21.5	316.4	2893.7	25.6
1991	126.7	355.3	26.7	359.2	3058.4	28.4
1992	156.3	424.4	33.6	407.5	3212.6	31.8
1993	190.2	471.3	55.9	439.0	3368.4	34.3
1994	220.8	486.3	73.2	460.7	3546.1	35.0
1995	252.9	513.2	107.7	459.0	3719.3	35.8
1996	286.5	548.2	137.8	473.1	3954.5	36.6
1997	316.6	579.1	164.5	476.0	4200.4	36.6
1998	415.4	637.1	255.7	478.4	4790.5	37.3
1999	523.1	679.1	322.9	537.9	5055.5	40.8
2000	607.7	706.7	385.5	576.1	5508.6	41.3
2001	706.3	863.4	503.8	653.1	6102.6	44.7
2002	820.6	1040.4	589.9	729.8	6896.3	46.1
2003	919.6	1300.5	660.5	752.2	7797.0	46.6
2004	925.2	1408.0	664.6	852.3	8872.5	43.4
2005	736.8	1598.9	709.5	974.2	10049.0	40.0
2006	726.4	1777.6	700.0	1122.8	11112.9	38.9
2007	723.6	2118.9	710.0	1381.9	11955.4	41.3
2008	768.0	2289.5	748.7	1402.7	11911.1	43.7
2009	745.3	2432.8	717.0	1495.3	11707.7	46.0

funds, etc.). As of year-end 2009, Fannie Mae had $2433 billion in its MBS outstanding in investors' hands; Freddie Mac had $1495 billion outstanding (see Table 1). The MBS represent a 'pass-through' claim on the streams of interest payments and principal repayments by the underlying mortgage borrowers. Since the investors might otherwise be leery of investing in such securities because of the unknown repayment prospects of the underlying borrowers, both Fannie and Freddie guarantee repayment to the investors, for which they have charged annual 'guarantee fees' (which are approximately 0.20–0.25%, or 20–25 basis points) on the unpaid principal and against which they are required to set aside a small amount of capital ($0.45 per $100 of guaranteed MBS).

One important feature of the GSEs' MBS is worth keeping in mind: although the GSEs' guarantees protect their MBS investors against the credit risk of their underlying borrowers' defaulting, the MBS investors are nevertheless exposed to interest-rate risk, since the underlying mortgages typically have a 30-year maturity. Further, because the mortgage borrowers can always pre-pay their mortgage principal without paying any fees (i.e. they can exercise their 'option' to pre-pay at no explicit cost to themselves at the time of exercise), the interest-rate risk that the MBS investors face is thereby heightened: when interest rates increase (above the contract rate on the mortgage), the MBS will be worth less to the investors (which is the standard risk that fixed-rate lenders face); but when interest rates decrease (below the contract rate), the mortgage borrowers are likely to pre-pay their mortgages and refinance at the new (lower) rates, thus depriving the investors of the capital gain that would normally occur on a fixed-rate instrument (and forcing the investors to have to reinvest their funds at the lower rates).

Why Fannie and Freddie Have Been Treated as Special

Prior to their conservatorships, both Fannie Mae and Freddie Mac might have appeared, at first glance, to be ordinary US corporations: Their corporate structures appeared quite ordinary, with chief executive officers (CEOs) and boards of directors, and their shares of stock could be bought and sold on the New York Stock Exchange.

However, there was much more to them, which differentiated them from other corporations and made them quite special (White 2003, 2004; Frame and White 2005):

- Their corporate charters were created through specific congressional legislation;
- The board of directors of each company was mandated to have 18 members, of which the President of the United States could appoint five members;
- They paid no state or local income taxes;
- They each had a potential line of credit with the US Treasury of up to $2.25 billion;
- Their securities were considered to be 'government securities' under the Securities Exchange Act of 1934;
- They were not required to register their securities with the US Securities and Exchange Commission (SEC), and they were exempt from SEC fees;
- Their securities could be purchased and held in unlimited quantities by US banks and savings institutions;
- Their securities could be purchased by the Federal Reserve for the latter's 'open market operations';
- They each could use the Federal Reserve as their fiscal agent; and
- Their insolvencies could not be resolved by a bankruptcy process or by a regulatory agency but instead would have to be resolved by the US Congress.

There were also limitations:

- Their activities were specifically restricted (again, by statute) to the secondary mortgage market; they were specifically prohibited from originating mortgages;
- The size of mortgage that they could buy (the 'conforming loan limit'), either for investment or for securitization, was limited in amount (which was adjusted each year in accordance with an index of house prices); as of early 2008 that amount was $417,000 (but Congress subsequently expanded this amount for high-cost housing areas to as high as $729,750);
- They were subject to prudential regulation by a federal regulatory agency (until 2008, this was the Office of Federal Housing Enterprise Oversight [OFHEO]; in the summer of 2008 the Federal Housing Finance Agency [FHFA] replaced OFHEO); and
- They were subject to 'mission regulation' (i.e. regulatory requirements that they meet targets with respect to their mortgage purchases in areas with low-and moderate-income and underserved households), which was under the jurisdiction of the US Department of Housing and Urban Development (HUD) until the summer of 2008 (when FHFA absorbed this role).

It was thus no accident that the GSE label came to be applied to these two companies.

There was one additional feature about the two companies that made them special: their sheer size. Their combined mortgage ownership and mortgage guarantees meant that they were involved with approximately $5 trillion in US residential mortgages (which, in turn, meant that they were involved with over 40% of the US residential mortgage market).

The GSEs' specialness had an important (and wholly intended) consequence: they could borrow at interest rates that were more favourable (i.e. lower) than their financial condition would otherwise justify. The consensus of academic studies was that this borrowing advantage was approximately two-fifths of a percentage point (40 basis points) (Frame and White 2005). In essence, the financial markets believed (correctly, as events turned out) that if either of the two companies were ever in financial difficulties, the US Treasury would very likely rescue ('bail out') their creditors – despite the explicit language that

accompanied all of their debt securities that these securities were not 'full faith and credit' obligations of the US Government. This belief on the part of the financial markets came to be known as the belief in the US Treasury's 'implicit guarantee'.

In turn, their favourable borrowing costs had a consequence for residential mortgages: Mortgages that were within the conforming loan limits carried interest rates that were approximately a quarter of a percentage point (25 basis points) lower than larger ('jumbo') mortgages that the GSEs were not permitted to buy (Frame and White 2005).

Until their insolvencies and conservatorships in 2008, the GSEs seemed to be providing a 'free lunch': they caused interest rates on conforming mortgages to be lower, without any apparent need for explicit budgetary subsidy from the federal fisc. It is not surprising that the GSEs enjoyed wide popularity in Congress.

The GSEs' Origins, and Subsequent Developments Through the 1980s

Fannie Mae began in 1938 as a federal agency, designed to buy and hold residential mortgages, using borrowed money (which, since it was a federal agency, meant US Treasury borrowings). In essence, this meant that Fannie Mae was channeling more funds into the mortgage market. Fannie Mae's operations were part of the larger efforts of President Franklin D. Roosevelt's New Deal to bring the US economy out of the Great Depression (including substantial efforts at assisting the housing sector through the mortgage insurance that was provided through the Federal Housing Administration [FHA]).

Through the 1950s and most of the 1960s Fannie Mae's growth was modest; as late as 1965, its mortgage holdings accounted for less than 2% of all residential mortgages in the USA (see Table 1). Most mortgages at the time were originated, and held in portfolio, by US savings and loan (S&L) institutions. Nevertheless, Fannie Mae had an important symbolic position as part of the federal government's efforts to assist housing.

Beginning in 1965 Fannie Mae grew more rapidly; and in 1968, as part of an effort to reduce the apparent size of the US Government's debt (and also because of a budgetary accounting quirk that would have meant that Fannie Mae's mortgage purchases would be scored as government expenditures and thus would contribute to the annual budget deficit), the Johnson administration privatized Fannie Mae (and its associated debt). In essence, Fannie Mae became a publicly traded company, but it retained the array of special features that were listed above and thus became a true GSE (although the term itself did not come into widespread use until the 1990s).

Fannie Mae was replaced within the US Department of Housing and Urban Development by the Government National Mortgage Association ('Ginnie Mae'), which was tasked with developing a method of securitizing the residential mortgages that were being insured by the FHA and by the US Veterans Administration (VA). The first Ginnie Mae MBS were issued in 1970.

The US S&L industry had largely shunned Fannie Mae, seeing it as the instrument of (and for) the non-depository mortgage finance companies (which have subsequently come to be known as 'mortgage bankers'). The S&Ls wanted a secondary mortgage market entity of 'their own'. Congress complied and created Freddie Mac in 1970, though it was initially owned by yet another GSE (the Federal Home Loan Bank System [FHLBS], which in turn was owned at the time by the S&L industry). Freddie Mac immediately began buying loans, and in 1971 Freddie Mac issued its first MBS. (Fannie Mae was slow to develop MBS and did not issue its first MBS until 1981.) Since Freddie Mac was owned by the FHLBS, and ultimately by the S&L industry, there was a certain logic to having the GSE governed by the federal regulator of the S&L industry and of the FHLBS (the Federal Home Loan Bank Board), which is what Congress arranged when it created Freddie Mac.

The US S&L industry hit hard times in the late 1970s, as accelerating inflation and then sharply higher interest rates made its basic model of accepting short-maturity deposits and lending these funds to borrowers for 30-year fixed-rate mortgage loans ('borrowing short and lending long') extremely problematic. Faced with a severe

interest rate squeeze in the high-interest rate environment, the S&L industry lobbied Congress for deregulation that would provide the industry with more flexibility, which Congress granted in 1980 and in 1982. Unfortunately, Congress neglected to increase the rigour of prudential regulation of the industry, which was needed to ensure that the new powers of flexibility would not be used for increased risk-taking (White 1991, 1993); instead, Congress did the opposite and weakened the prudential regulation of the industry. Hundreds of S&Ls took advantage of their new powers to take on enhanced risks, especially in the 'sunbelt' and 'oil patch' states of Florida, Louisiana, Texas, Arizona and California in the years 1983–1985. In the wake of decreasing prices of petroleum from 1981 through 1986 and changes in the US tax code in 1986 that made commercial real estate (in which these S&Ls had invested heavily) less attractive, these risky S&Ls failed, causing the industry to shrink.

Fannie Mae, which had a similar financial structure to that of the S&Ls (i.e. Fannie also borrowed short and lent long), suffered through a similar financial squeeze in the late 1970s and early 1980s. Although its accounting results continued to show that Fannie Mae was solvent, it was well known in Washington policy circles that the GSE was insolvent on a mark-to-market (i.e. market value) basis.Receding interest rates after 1982 allowed Fannie Mae to regain solvency later in the decade, even on a mark-to-market basis. But its 'near-death' experience chastened its senior management and limited its growth in on-balance-sheet mortgages for the remainder of the 1980s, though it did expand its MBS business.

The GSEs' Growth and Further Developments in the 1990s and 2000s

By the beginning of the 1990s, Fannie Mae and Freddie Mac were ready to expand substantially and replace the ailing and shrunken S&L industry as the dominant influence in residential mortgage finance. Fannie Mae finally shook off its trauma of the early 1980s. Freddie Mac, which had been somewhat restrained by its federal governors (the Federal Home Loan Bank Board) in the 1980s, at least in terms of its on-balance-sheet mortgage assets, was converted into a publicly traded company (but, again, with the full array of special features listed above) by Congressional legislation in 1989 and thus was (like Fannie Mae) now a full-fledged GSE. With a board of directors that was now answerable to shareholders, and with shareholders eager for the profits that could come from rapid growth, Freddie Mac (like Fannie Mae) was ready to grow.

The growth experience of both GSEs is shown in Table 1. It is clear from the table that both GSEs grew rapidly, in terms of the mortgage assets that were on their balance sheets and the mortgages that they were converting into MBS, from 1990 through 2000 and also from 2000 through 2003. There are multiple reasons for their growth:

- The decline of the S&L industry left a gap in the residential mortgage finance area;
- Both GSEs were primed for growth after the restraints of the 1980s;
- The process of securitization as a new technology for mortgage finance did offer efficiencies compared with the 'traditional' process of mortgage finance through depository institutions;
- The two companies' status as GSEs gave them the borrowing advantage that was described above, making it advantageous for them to expand through borrowing;
- When they bought mortgages and held those mortgages in their own portfolios, the GSEs were required to hold only 2.50% capital against those mortgages; by contrast, S&Ls and commercial banks (which were also trying to expand to fill the gap that was left by the shrinking S&L industry) were required to hold 4.00% capital against mortgages that they held, so the GSEs had a clear cost advantage (since equity is generally more costly than is debt) in holding mortgages; and
- When they bought mortgages and converted them into guaranteed MBS, the GSEs were required to hold only 0.45% capital against the guarantees; when banks or S&Ls bought these GSE MBS as investments, these depositories needed to hold only 1.60% in capital

against these investments (as compared to the 4.00% in capital that they needed when they held the original 'whole loan' mortgages); thus, even though they had to pay an annual guarantee fee to the GSEs, depositories generally found it worthwhile to swap their whole loan mortgages for GSE MBS, which fuelled the rise of the GSEs' MBS business.

Critics of the GSEs in the late 1990s and early 2000s worried that the combination of the large (and rapidly growing) sizes of their on-balance-sheet assets, which were primarily 30-year fixed-rate mortgages, and their thin capital levels – recall that the liabilities side of their balance sheets had 96–97% debt and only 3–4% equity capital – meant that the two companies were exposed to excessive interest-rate risk that could cause their insolvencies (Wallison 2000, 2001; Jaffee 2003; White 2003, 2004; Frame and White 2005). The two companies' public assurances that they were adequately containing their interest-rate risk through the use of derivatives did little to reassure the critics, since the details of the derivatives activities were not public information.

The asset growth paths of the two companies came to a halt around 2003–2004 because of accounting scandals that first engulfed Freddie Mac (in 2003) and then Fannie Mae (in 2004). The two companies' accounting irregularities provided their prudential regulator (OFHEO) with sufficient leverage to claim that they were operating in an 'unsafe and unsound' condition and that their asset growth needed to be contained (and they needed to maintain higher levels of capital). Since their MBS issuances did not create interest-rate risk for the GSEs, these activities were not restrained (and most critics did not complain).

What was overlooked by the critics at the time was the deteriorating quality of the mortgages that Fannie and Freddie were buying (Acharya et al. 2011, Chs. 2–3).

In order to limit the credit risk to which they might otherwise be exposed (which was especially important because of their thin capital levels), the GSEs were supposed to buy only high-quality mortgages that met 'investment quality standards' (as determined by OFHEO, their prudential regulator). In the early 1990s and before, these had usually meant mortgage loans where the borrower had made at least a 20% down payment (or, equivalently, the loan-to-value [LTV] ratio was 80% or less) or had private mortgage insurance for loans where the down payment was as little as 5%; where the borrower had a good credit history (as represented by a good 'credit score' that was usually compiled by Fair, Isaac and Company and that came to be known as the 'FICO score'); and where the borrower's income was deemed adequate so that the monthly payments on the mortgage were affordable. These indicia meant that the borrower was unlikely to default and that even in the event of default the sizable down payment (or mortgage insurance) provided a buffer that would protect the GSEs (as investor or as guarantor) against loss.

But, beginning in the mid-1990s, these credit quality standards began to slip (Acharya et al. 2011, Ch. 2) – partly because lower-quality mortgages provided an additional path for expansion for the GSEs and partly because the regulatory pressures on the GSEs to expand their mortgage purchases from low-and moderate-income households and households that were located in underserved areas were increasing. The general upward trend in housing prices in the USA, which especially picked up steam around 1996, masked this deterioration. In an environment where housing prices are generally rising, the standard quality indicia become less important: even if the borrower experiences an adverse shock – she is involved in a severe accident or otherwise becomes unemployed – and thereby cannot make the monthly payments on her mortgage, she can avoid defaulting on that mortgage by selling the house (at a profit) and paying off the mortgage through that route. Indeed, the GSEs experienced credit losses on their combined mortgage holdings plus MBS outstanding that were annually below 0.1% (!) from 1996 onward (FHFA 2010).

The GSEs' involvement in lower quality mortgages became substantially greater around 2003 (Jaffee 2010; Acharya et al. 2011, Ch. 3). From 2000 onward, the growth in sub-prime mortgage lending and securitization threatened the market shares of the GSEs. At first glance, this should not

have been so, since the high quality standards of the GSEs should have kept them separated and aloof from the sub-prime lenders and borrowers, and vice versa. However, in the environment of rising housing prices, mortgage borrowers who otherwise would have qualified for a conforming loan were being encouraged by lenders to borrow larger amounts (which would push them into 'jumbo' territory) and/or to structure their loans in ways that would not meet the GSEs' underwriting standards (which would push them into non-conforming territory). The latter was done, for example, by the borrower's making less than the requisite 20% down payment but not arranging for (costly) private mortgage insurance, or by getting a second mortgage loan to cover some or even all of the down payment, or by getting an initial low 'teaser' interest rate but with a scheduled upward adjustment after two or three years.

Supplementing these market-share pressures were the aforementioned regulatory pressures to expand the GSEs' purchases of mortgages from low-and moderate-income households and from households in underserved areas. These regulatory pressures also led to the GSEs' decisions to purchase significant volumes of 'private label' AAA-rated MBS (i.e. MBS that were issued by banks and other issuers that were not GSEs) that had sub-prime mortgages as their underlying collateral, since many of these sub-prime borrowers were households in the designated categories and the GSEs received regulatory credit for these securities purchases.

Again, rising home prices initially masked the consequences of these actions. Credit losses at both GSEs remained well below 0.05% from 1999 through 2006. But the Case-Shiller national index of home prices peaked in the second quarter of 2006 and then began to decline. Without the 'you-can-surely-sell-the-house-at-a-profit' safety valve for borrower difficulties, mortgage delinquencies began to rise, and then defaults followed. The increases were especially sharp for sub-prime mortgages, but all categories of mortgages suffered increases, including (not surprisingly) GSE mortgages. The pattern of cumulative defaults by year of origination can be seen in Fig. 1 for Fannie Mae and Fig. 2 for Freddie Mac. It is clear that 2004 marked the beginning of a different default experience, due to a combination of the lower quality of the mortgages that the GSEs bought and the lesser amount of time for house price appreciation to cover the sins of lower quality. The successive annual cohorts were appreciably worse.

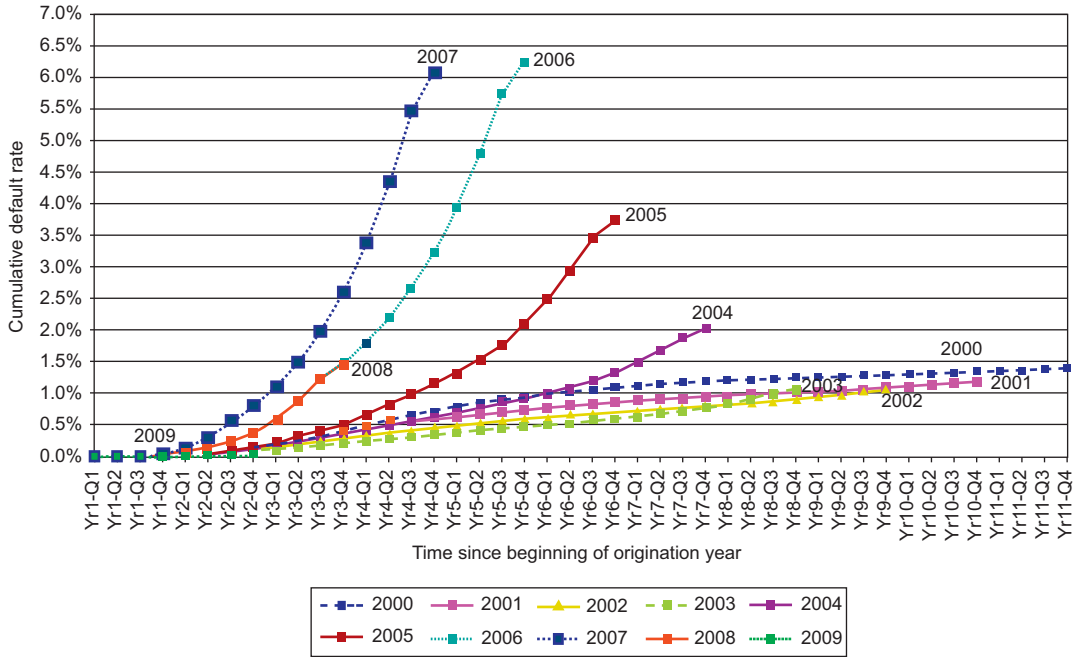
The rising defaults on sub-prime mortgages and then on the MBS that were based on sub-prime mortgages also meant that the GSEs' experienced losses on their investments in those apparently safe AAA-rated private label MBS.

The GSEs failed to earn profits in 2007, instead running losses – for the first time ever for Freddie Mac, and for the first time since 1985 for Fannie Mae.

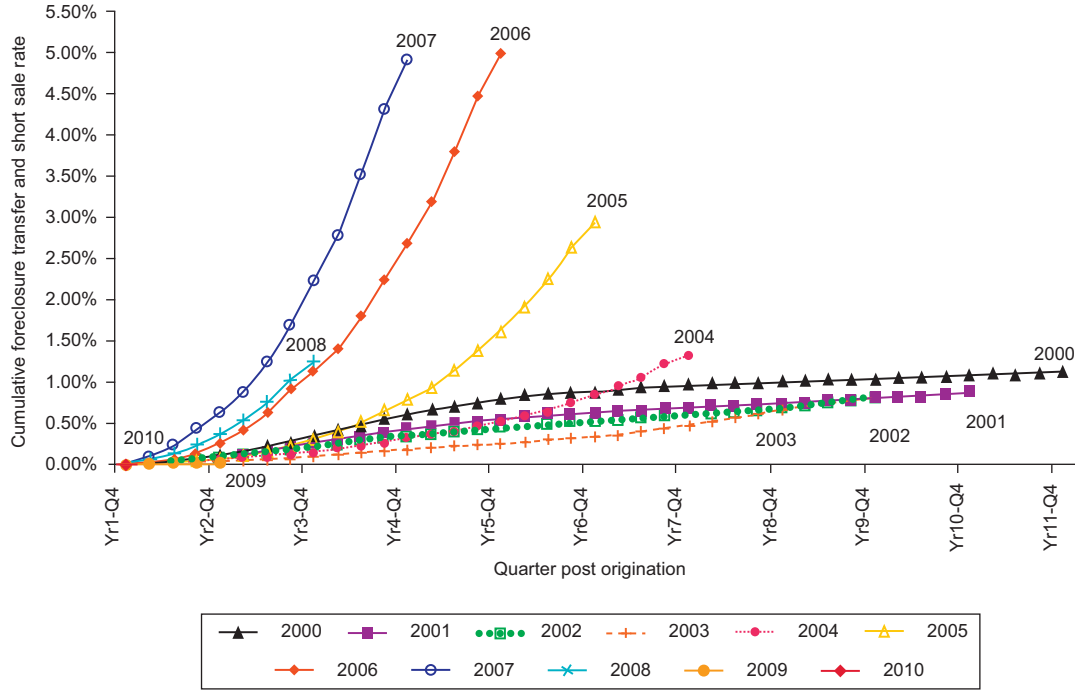
In the first two quarters of 2008, the losses for both GSEs continued to rise. Although the delinquencies on GSE mortgages were at lower rates than for other mortgages (see, for example, Fannie Mae 2011, p. 13, and Freddie Mac 2011, p. 17), nevertheless their thin capital levels were an insufficient buffer against these losses. By the end of the summer of 2008, their insolvencies were looming, and the capital markets began to worry whether the Treasury really would come to the rescue. As Jaffee (2010) points out, although the deteriorating credit quality of the GSEs' mortgages was the ultimate problem, the immediate problem that the GSEs faced was their difficulties in rolling over their short-term debt – in financing themselves. On 6 September 2008, in coordination with the Treasury, the FHFA placed them into conservatorships. In principle, the companies were still intact, with their shareholder/owners still in place; in practice, they had become wards of the US Government (which immediately dismissed and replaced their senior managers). The Treasury agreed to cover their losses and thus keep their creditors whole. (Accounts of the Treasury's day-to-day and hour-to-hour decisions can be found in Paulson (2010) and Sorkin (2009)). The financial markets' belief in the implicit guarantee had proved correct.

Epilogue

As of the spring of 2011, Fannie Mae and Freddie Mac remain in conservatorships, but also remain



Fannie Mae, Freddie Mac and the Crisis in US Mortgage Finance, Fig. 1 Fannie Mae cumulative default rates by year of origination



Fannie Mae, Freddie Mac and the Crisis in US Mortgage Finance, Fig. 2 Freddie Mac cumulative default rates by year of origination

key participants in the secondary mortgage market. The Treasury has had to make capital contributions of over $150 billion in the two GSEs to cover their accumulated losses. Because there are still pre-2008 mortgages that have not yet defaulted but that are highly likely to do so over the next few years, the GSEs' accumulated losses will probably increase, to at least $200 billion, and possibly as high as $400 billion. It is these delayed-recognition losses that continue to make headlines every quarter as the GSEs announce their latest financial results.

Because of the trauma of the collapse of the sub-prime mortgage securities market and of the financial markets more generally, residential mortgage lending remains in a fragile condition, with Fannie and Freddie purchases and guaranteed MBS accounting for about 70% of mortgage originations and FHA guarantees (and Ginnie Mae MBS) accounting for about 20%. Government guarantees thus are involved with over 90% of mortgage originations. The GSEs appear to have tightened their quality standards, back to their pre-1990s levels (with 20% down payments, etc.); the early (lower) foreclosure results for the 2009 are consistent with this claim (see Figs. 1 and 2). Thus, their new mortgage activity is unlikely to exacerbate their losses.

Nevertheless, there is a general consensus that their quasi-private, quasi-public status is highly problematic (Acharya et al. 2011) and that they should be phased out, with an expanded private presence to replace them. But what, if any, federal government role should persist in the general residential mortgage market remains an open question. The Obama administration formulated its proposals in February 2011. Congress has yet to act. The GSEs' current status in conservatorship limbo could well endure for a few years before a political consensus on their phase-out (and what would replace them) is reached.

See Also

- ▶ Credit Crunch Chronology: April 2007–September 2009
- ▶ Run on Northern Rock

Bibliography

Acharya, V.V., M. Richardson, S. Van Nieuweburgh, and L.J. White. 2011. *Guaranteed to fail: Fannie Mae, Freddie Mac, and the debacle of mortgage finance*. Princeton: Princeton University Press.

Fannie Mae. 2011. *Fannie Mae 2010 credit supplement*, 24 February 2011. Available at: http://www.fanniemae.com/media/pdf/newsreleases/q42010_credit_summary.pdf; jsessionid = JTHEGVQCJCC1DJ2FQSHSFGA

Federal Housing Finance Agency. 2010. *Report to Congress*, 2009.

Frame, W.S., and L.J. White. 2005. Fussing and fuming over Fannie and Freddie: How much smoke, how much fire? *Journal of Economic Perspectives* 19(2): 159–184.

Freddie Mac. 2011. *Fourth quarter 2010 financial results supplement*, 24 February 2011. Available at: http://www.freddiemac.com/investors/er/pdf/supplement_022411.pdf

Jaffee, D.M. 2003. The interest rate risk of Fannie Mae and Freddie Mac. *Journal of Financial Services Research* 24(1): 5–29.

Jaffee, D. M. 2010. *The role of the GSEs and housing policy in the financial crisis*. Prepared for the Financial Crisis Inquiry Commission, 27 February.

Paulson, H.M. 2010. *On the brink: Inside the race to stop the collapse of the global financial system*. New York: Business Plus.

Sorkin, A.R. 2009. *Too big to fail*. New York: Viking.

Wallison, P.J., eds. 2000. *Public purposes and private interests: Fannie Mae and Freddie Mac*. Washington, DC: American Enterprise Institute.

Wallison, P.J., eds. 2001. *Serving two masters yet out of control: Fannie Mae and Freddie Mac*. Washington, DC: American Enterprise Institute.

White, L.J. 1991. *The S&L debacle: Public policy lessons for bank and thrift regulation*. New York: Oxford University Press.

White, L.J. 1993. A cautionary tale of deregulation gone awry: The S&L debacle. *Southern Economic Journal* 59(3): 496–514.

White, L.J. 2003. Focusing on Fannie and Freddie: The dilemmas of reforming housing finance. *Journal of Financial Services Research* 23(1): 43–58.

White, L.J. 2004. Fannie Mae, Freddie Mac, and housing finance: Why true privatization is good public policy. *Policy Analysis*, No. 528, Cato Institute, 7 October.

Fanno, Marco (1878–1965)

Joseph Halevi

Fanno was a most distinguished Italian economist who became Professor of Political Economy in

1909 and taught at the universities of Sassari, Cagliari, Messina and Padua.

His work places him between the Italian tradition of General Equilibrium and the macrodynamic theories developed during the 1930s. From this perspective, Fanno was unique among the scholars who shaped Italy's economic thought until the end of World War II. Indeed, most economists were reared in the General Equilibrium school of Pareto and Pantaleoni and did not absorb the new formulations of the 1930s.

Fanno's contributions range from the theory of joint costs (1914) to the analysis of the elasticity of demand (1929, 1933) and monetary issues (1913, 1937). Yet it is a study on economic fluctuations that constitutes Fanno's most important work (1947). This study is characterized by a systematic sifting of the major theoretical literature on the subject, as well as of a large amount of historical and empirical material. Analytically, his approach to the trade cycle reflects Ragnar Frisch's model of the propagation of impulses in economic activity. In his book, Fanno discusses in detail the role of credit in determining the duration of the cycle. In this respect he departed from the theories of the real trade cycle and moved closer to Keynes's *Treatise on Money.*

Selected Works

1913. *Le banche e il mercato monetario*. Rome: Loescher.
1914. Contributo alla teoria dell'offerta a costi congiunti. *Giornale degli Economisti* 49-(Supplement): 1–143.
1929. Die Elastizität der Nachfrage nach Ersatzgütern. *Zeitschrift für Nationalökonomie* 1(1): 51–74.
1933. Interrelation des prix et courbes statistiques de demande et d'offre. *Econometrica* 1(2): 162–71.
1937. *Lezioni di economia e legislazione bancaria*. Padua.
1947. *La teoria delle fluttuazioni economiche*. Turin: Unione Tipografico-Editrice Torinese.

Farr, William (1807–1883)

R. M. Smith

Keywords
Farr, W.; Human capital; Mortality

JEL Classifications
Q1

William Farr, born in Kenley, Shropshire on 30 November 1807, died in London on 14 April 1883, was a statistician in the General Register Office who had been appointed in 1840 as 'compiler of abstracts' and was two years later made Statistical Superintendent, a post he held until his retirement in 1880. He pioneered the quantitative study of morbidity and mortality and in the process became one of Victorian England's most prominent figures in the public health and reform movements (Cullen 1975). He made major contributions in the fields of data collection, being largely responsible for the introduction of a cause of death classification which was linked with his derivation of the 'zymotic' theory of epidemic disease (Eyler 1979; Pelling 1978). As an Assistant Census Commissioner for each of the censuses of 1851, 1861 and 1871, he was largely responsible for the development of reliable procedures for the recording of occupations (McDowall 1983). He is, however, best known as a statistical analyst, for in 1843 he constructed the first English Life Table based on deaths in 1841 linked to the census of that year. At the same time he established the formula for deriving from a rate of mortality by age m the probability of survival p at the initial age. In 1850 and 1864 Farr produced his second and third English Life Tables, the last mentioned being used as the actuarial basis for the life insurance scheme set up by the Post Office for its employees. Farr in his work on occupational mortality was the first to make extensive use of the standard mortality rate, allowing comparisons of the mortality of different groups by means of a summary statistic which took account of differences in the age structure of

the groups being compared. A recurring theme in his work was the identification of variation in mortality in different urban areas of the country. Such differential mortality was viewed as an index of human welfare. For example, in 1850 one-tenth of the registration districts, those he named 'healthy districts', had average mortality rates not exceeding 17 per 1,000, a rate he thought indicative of the 'natural' mortality which, when exceeded, would indicate those deaths attributable to unnatural and preventable diseases. An underlying aim in much of his work was to discover statistical laws or numerical expressions of regularities such as he proposed in the laws of recovery and death in smallpox, the elevation law for cholera mortality in London (Lewes 1983) and the law of the relation between population density and mortality. He was also an early contributor to human-capital theory (Kiker 1968) arguing, in particular, that the economic value of men varied with age as well as social class, and this he used as powerful publicity for urban reform by drawing attention to the financial losses that followed from diseases that were the causes of death and illness in society at large.

Selected Works

1843a. Causes of the high mortality in town districts. *5th annual report of the registrar general,* 406–435 (*Parliamentary Papers* XXI, 200–215).

1843b. English life Table no. 1. *5th annual report of the registrar general,* 354–358, 366–367 *(Parliamentary Papers* XXI, 168–171, 178) and *6th annual report of registrar general,* 517–666 (*Parliamentary Papers,* 1844, XIX, 290–358).

1850. English life Table no. 2: Males. *12th annual report of the registrar general,* Appendix, 73–152.

1852a. Influence of elevation on the fatality of cholera. *Journal of the Statistical Society* 15, 155–183.

1852b. *Report on the mortality of Cholera in England, 1848–49.* London: HMSO.

1854. Vital statistics. In *A descriptive and statistical account of the British Empire: exhibiting its extent, physical capacities, population, industry, and civil and religious institutions,* ed. J.R. McCulloch, 4th edn. London: Longman, Brown, Green and Longmans.

1859. English life Table no. 2: Females. *20th annual report of the registrar general,* Appendix, 177–203 (*Parliamentary Papers,* 1859, sess. 2 XII).

1864. English life Table: *Tables of lifetimes, annuities and premiums, with an introduction by William Farr, M.D., F.R.S., D.C.L. London.*

1866. Mortality of children in the principal states of Europe. *Journal of the Statistical Society* 29, 1–35.

1867–8. *Report of the Cholera Epidemic of 1866 in England: supplement 29th annual report on the registrar general* (*Parliamentary Papers* XXXVI).

1885. *Vital statistics: A memorial volume of selections from the reports and writings of William Farr, M.D. D.C.L. C.B. F.R.S,* ed. N.A. Humphreys. London.

Bibliography

Cullen, M.J. 1975. *The statistical movement in early Victorian Britain: The foundations of empirical social research*. Brighton: Harvester Press.

Eyler, J.M. 1979. *Victorian social medicine: The ideas and methods of William Farr.* Baltimore/London: Johns Hopkins University Press.

Kiker, B.F. 1968. *Human capital: In retrospect*, Essays in economics no. 16. Columbia: Bureau of Business and Economic Research, University of South Carolina.

Lewes, F. 1983. William Farr and cholera. *Population Trends* 31 (Spring): 8–12.

McDowall, W. 1983. William Farr and the study of occupational mortality. *Population Trends* 31 (Spring): 21–24.

Pelling, M. 1978. *Cholera, fever and English medicine*. Oxford: Oxford University Press.

Farrell, Michael James (1926–1975)

Christopher Bliss

Keywords

Andrews, P.; Community indifference curves; Competitive equilibrium; convexity; Farrell,

M. J.; Firm; Theory of; Linear programming; Profit maximization; Rational expectations; Stone, J. R. N.; Subjective probability; Yield gap

JEL Classifications
B31

M.J. Farrell was born in 1926 and read Politics, Philosophy and Economics at New College, Oxford, graduating with First Class Honours. He moved to Cambridge in 1949 to work with Richard Stone at the Department of Applied Economics. He became a Fellow of Gonville and Caius College and the University made him Lecturer in Economics and eventually Reader. He was Editor of the *Review of Economic Studies* and a Fellow of the Econometric Society. In 1957 Farrell contracted poliomyelitis which left him dependent on crutches to get about. He died in 1975.

The bibliography of Farrell's work provided by Fisher (1976) lists 25 journal papers, about one a year in a cruelly shortened academic life. The quality of these papers is remarkable. They reveal the clarity of their author's mind and an outstanding creativity. Farrell often answered questions that others had hardly considered.

As a young man Farrell was influenced by Phillip Andrews, the author of *Manufacturing Business,* and they shared a dissatisfaction concerning the prevailing theory of the firm: 'They [economists in the 1920s and 1930s] reduced the theory of the firm to a maximization problem soluble by the most elementary application of the differential calculus ...' and 'Unfortunately these conclusions did not fit the regrettably complex facts well ...' (Farrell 1971, p. 10). Farrell's work on the theory of the firm displayed an acute understanding of the subtlety of profit maximization as a strategy. In (1954) he provided one of the first applications of linear programming to this field. Farrell believed that the case for profit maximization eventually depended in part on the operation of a selection process. His (1970) paper remains to this day one of the best papers ever written on that topic.

Farrell wrote on the measurement of productive efficiency, on the consumption function, and on welfare economics. On some topics he produced a single paper – his last was on social choice theory.

In (1959) Farrell made two observations which were important innovations at the time. First, he exposed what he called 'the fallacy'. This is a confusion between sufficient and necessary conditions for competitive equilibrium to be efficient. Convexity, as Farrell neatly demonstrated, is *sufficient* for existence of equilibrium but is *necessary* neither for existence nor for efficiency. Second, '... concavities in individual indifference maps disappear when one aggregates over a large enough number of individuals' (1959, p. 381).

This deep aggregation result, which gave rise to an extensive literature (see, for example, Arrow and Hahn 1971, chs 7 and 8), is based on a simple point. To illustrate it consider consumers and let them all have the same tastes, which may be represented by $U(x)$, where x is a vector of consumptions. Suppose that $U(x_1) = U(x_2)$ and let there be N consumers. We now wish to see whether a convex combination of $N \cdot x_1$ and $N \cdot x_2$, that is $\lambda \cdot N \cdot x_1 + (1 - \lambda) \cdot N \cdot x_2$, can be distributed so as to make each consumer at least as well off as with x_1 or x_2. If it can, community indifference curves will be convex even if those derived from $U(\ )$ are not.

If consumers were indefinitely divisible we could achieve this result by giving x_1 to $\lambda \cdot N$ consumers and x_2 to $(1 - \lambda) \cdot N$ consumers. However, as $\lambda \cdot N$ may not be an integer this exact procedure is inadmissable. Nevertheless, as N becomes large an integer $M < N$ will eventually emerge such that M/N approximates λ to any desired degree of accuracy. Hence Farrell's result follows.

Farrell treated the often sloppily discussed question whether speculation could be destabilizing and still profitable, in (1966). His demonstration within a very general framework that linearity of demand functions is required to exclude this possibility greatly advanced the general understanding of this problem.

In (1962) Farrell considered the well-known problem of the yield gap, the observation that equities at certain times show a different rate

of return from that obtained from bonds. He provided some calculations which showed that there had been yield gaps in the past even when returns were corrected for capital gains. In considering what light these *ex post* observations throw on investors' *ex ante* decisions, Farrell asked '... what do we mean by perfect knowledge in a market where uncertainty is present?' (1962, p. 835). This led him to analyse what he called 'accurate' expectations: '... an individual's expectation is "accurate" if his subjective probability distribution is the same as the hypothetical frequency distribution by which we represent the real world' (1962, p. 836). Long before the idea of rational expectations became fashionable, Farrell saw its relevance to the analysis of securities markets. However the careful student of profit maximization and selection processes found no reason to assume that expectations would necessarily be 'accurate'.

Selected Works

1954. An application of activity analysis to the theory of the firm. *Econometrica* 22: 291–302.
1959. The convexity assumption in the theory of competitive markets. *Journal of Political Economy* 67: 377–391.
1962. On the structure of the capital market. *Economic Journal* 72: 830–844.
1966. Profitable speculation. *Economica* 33: 183–193.
1970. Some elementary selection processes in economics. *Review of Economic Studies* 37: 305–319.
1971. Philip Andrews and manufacturing business. *Journal of Industrial Economics* 20: 10–13.

Bibliography

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. Amsterdam: North- Holland.
Fisher, M.R. 1976. The economic contribution of Michael James Farrell. *Review of Economic Studies* 43: 371–382. (Includes a complete discussion of Farrell's works.)

Fascism

Wolfgang-Dieter Classen

The term fascism can be applied to historical reality only as an approximation, because the differences between what are called fascist movements and regimes seem to be greater than the similarities, and leave room for many contrary interpretations (cf. de Felice 1969; Gregor 1974). Given this restriction the term is applied to both radical populistic mass movements, primarily of the middle classes, and, where they attained power, to the political regimes they created between the two world wars.

The fascist movements emerged as a result of the political, economic and social crisis of the bourgeois societies in European countries after World War I. They propagated an extreme anti-liberal, anti-socialist, nationalist and imperialist (and, in Germany, racial) ideology, and above all, they struggled with militancy and terror against the labour organizations. Where these movements came to power (Italy and Germany) it was by coalition with the bourgeois upper class and thanks to the simultaneous failure of labour organizations to present any effective resistance. The political structure of the fascist regimes was, on the surface, marked by the dictatorial leader, the single party system, the total control of the press and all information sources, massive propaganda campaigns, tendencies toward the coordination of all political, economical, social and cultural institutions from above, and the power of the party militia, the police and the secret police. But behind this surface of strictly hierarchical dictatorship the fascist leaders' disregard for administration, their glorification of struggle and competition as an ideological expression of Social Darwinism led to a lack of constitutionality, to a deficient division of spheres of control and influence between the agencies, and, especially in the later years, to a multiplication of hurriedly erected ad hoc Commissariats without any proper plan of coordination. That, in turn, left much room

for constant quarrels and boundary disputes between the party leaders, representatives of special party organizations (e.g. the SS, the Arbeitsfront in Germany), the army, the state machinery (traditionally the realm of the conservative bourgeoisie) and big industry as rival power blocs. This disintegration of the regime's power structure often made political decision procedures very ineffective. (With regard to Germany, see Fraenkel 1941; Neumann 1944; Broszat 1969; Hirschfeld and Mommsen 1980.)

Fascism and the Economy

Fascism did not lead to any original contributions to economic theory except for some elements in the theory of corporatism added by Italian fascists. Positing the primacy of national over individual welfare, the fascist state was to direct economic activities for these purposes. In principle national interests meant economic strength on the basis of private ownership of the means of production, military power as a precondition for imperialistic expansion, independence in the world and autarky. These objectives implied in turn the necessity of rearmament. Thus in fascism the economy became ultimately an instrument of rearmament and autarky objectives; in Germany soon after fascism came to power (1934–5), in Italy during the World Depression that followed a period of relatively liberal economic policy (until 1926–7), in which a free-trade and a deflationary fiscal policy (to balance the budget) was implemented.

To revive the economy after the Depression the fascist regimes utilized deficit-financed government expenditures partly for infrastructural investments (like the Autobahnbau in Germany) but mainly for rearmament. Thus in Germany the total government expenditures as a proportion of gross national product doubled from 1932 to 1938. The armament expenditures as a proportion of GNP rose in the same time from nearly 1 per cent to more than 15 per cent, which in 1938 was 50 per cent of total government expenditure (Erbe 1958). In addition the regimes tried to stimulate civil economic activities – such as house renovation – by tax reductions and/or pecuniary aid.

Credit policy basically functioned as a means to finance the budget deficit. Because the public debt could not be totally financed from the private capital market, the credit institutions were obliged to absorb the public debt by accepting public treasury certificates. Thus the credit institutions lost their usual function as intermediaries in the private circulation of capital. They served instead as a collecting box of money to cover public debts. Tax credit notes and, in Germany, the so-called Mefo-bills were further financing instruments. The German Reich's debt increased from RM14 milliard in 1933 to RM42 milliard in 1938, of which RM12 milliard were raised by the Mefo-bills, showing the high proportion of short-term debts. As long as full employment had not been achieved this credit expansion had little inflationary effect.

The control over the volume of investment by prohibiting the distribution of dividends above a fixed level (in Germany, six per cent, by subjecting new issues of shares to the permission of the state and by obliging firms to lend the government all their non-invested excess capital were supportive measures to the management of deficit spending.

Falling imports and exports as a result of the Depression and the protectionism of the time led, especially in the fascist countries, to serious tendencies towards an insulation from cyclical trade movements and the creation of a closed economy. A neomercantilistic foreign trade policy became a means of achieving these objectives. Bilaterization of foreign trade, based on clearing and barter agreements accompanied by the use of economic, political and, later, military pressure to attain favourable trade arrangements; import licences; export subsidies; fixing of quotas; control over foreign exchange and high tariff barriers: all these instruments were used to regulate foreign trade totally with regard to the programmes of autarky and rearmament.

Thus, in accordance with the old imperialist aims of big business and as a preliminary to creating the closed *'Grossraumwirtschaft'*, German foreign trade shifted from the western to the weak southeast European countries with their large resources of raw materials (Sohn-Rethel 1973). The volume of German foreign trade with these

countries as a proportion of total German foreign trade more than doubled between 1932 and 1938. To get special raw materials German foreign trade with Latin America and northeast European countries developed in the same direction.

Based on growing internal demand Germany experienced rapid economic revival. Full employment had been achieved by 1937–8 from a situation of over six million jobless in 1932–3. Although this success served to establish mass loyalty toward the fascist regime, economic development was undoubtedly more for the benefit of the propertied classes and, above all, of big industry, whose profits in 1938 were twice as high as in 1932 (Bettelheim 1971, p. 232). As a result of the brutal destruction of all traditional independent labour organizations, the prohibition of strikes and the elimination of free wage negotiations, the degree of working class exploitation was increased, scarcely masked by some welfare services. While in Germany wages were fixed at the low level of the Depression year 1932, in Italy they were even cut. In Germany, the index of average weekly real wages reached the level of 1928 only in 1938, yet the average weekly labour time increased from 41.5 hours in 1932 to nearly 47 hours in 1938. Thus the growth of wages is to be seen as the result of rising working hours (Mason 1977, p. 149). Wages and salaries as a proportion of national income fell from 64 per cent in 1932 to 57 per cent in 1938.

The growing profits were mostly ploughed back into investments. In Germany the gross investment as a proportion of GNP rose from 9 per cent in 1932 to more than 15 per cent in 1938. Although personal consumption increased, total consumption as a proportion of GNP fell from 81 per cent in 1932 to less than 64 per cent in 1938 (Mason 1977, p. 149). The transformation of the production structure from consumer good industries to those of capital equipment was completely in line with the rearmament programme.

In pursuit of autarky, surrogates for imports and foreign raw materials were increasingly produced, shifting the orientation of many firms' production processes from the world to the domestic market. This often led to a loss of strong world market positions. This process was supported by a cartellization policy which was in contrast to the earlier anti-capitalist slogans of the fascist movement. Moreover, state-run factories were built up to increase the use of low-quality domestic raw materials with correspondingly high production costs. However, self-sufficiency could never be achieved. At the outbreak of the war Germany was still dependent on foreign supplies of oil, iron ore, manganese and many other raw materials (Kaldor 1945, p. 42).

With the intensification of measures for rearmament and autarky, after full employment had been achieved, beginning in Germany with the declaration of Hitler's 'Vierjahresplan' in 1936, public finances drifted towards a ruinous situation. Inflation was only suppressed by extensive controls of prices and wages. In an attempt to manage critical shortages of raw materials, quota systems were introduced. For the same reason, the employment of the labour force was increasingly controlled and directed. However, these interventions into the running of the economy took place without any proper planning.

Although the outbreak of the war necessitated the further intensification of armaments production German war potential was never fully exploited (Kaldor 1945). This would have meant the further extension of the average labour time, the employment of more women, the further reduction of consumer good production to the advantage of war production, and total planned economy. The reason the fascist leaders did not force the people to greater sacrifices is to be seen in their interpretation of Germany's defeat in World War I as a result of internal political instability (Mason 1977).

See Also

- ▶ Corporatism
- ▶ War Economy

Bibliography

Bettelheim, C. 1971. *L'économie allemande sous le nazisme. Un aspect de la décadence du capitalisme*. Paris: Maspero.

Broszat, M. 1969. *Der Staat Hitlers. Grundlegung und Entwicklung seiner inneren Verfassung*. Munich: DTV.

De Felice, R. 1966. Mussolini il fascista. I: La conquista del potere, 1921–1925, Turin: Einaudi, 1966; II: L'organizzazione dello Stato fascista, 1925–1929, Turin: Einaudi, 1968.
De Felice, R. 1969. *Le interpretazioni del Fascismo*. Bari: Laterza.
Erbe, R. 1958. *Die nationalsozialistische Wirtschaftspolitik 1933–1939 im Lichte der modernen Theorie*. Zurich: Polygraph Verlag.
Fraenkel, E. 1941. *The dual state. A contribution to the theory of dictatorship*. New York/London/Toronto: Octagon Books.
Gregor, A.J. 1974. *Interpretations of fascism*. Morristown: General Learning Press.
Hirschfeld, G., and W.J. Mommsen (eds.). 1980. *Der Führerstaat: Mythos und Realität. Studien zur Struktur und Politik des Dritten Reiches*. Stuttgart: Klett-Verlag.
Kaldor, N. 1945. The German war economy. *Review of Economic Studies* 13(1): 33–52.
Lyttleton, A. 1973. *The seizure of power: Fascism in Italy 1919–1929*. London: Weidenfeld &Nicolson.
Mason, T.W. 1968. The primacy of politics: Politics and economics in National Socialist Germany. In *The nature of fascism*, ed. S.J. Woolf. London: Weidenfeld & Nicolson.
Mason, T.W. 1977. *Sozialpolitik im Dritten Reich. Arbeiterklasse und Volksgemeinschaft*. Opladen: Westdeutscher Verlag.
Milward, A.S. 1965. *The German economy at war*. London: Athlone Press.
Neumann, F. 1944. *Behemoth. The structure and practice of national socialism*, 2nd ed. New York: Octagon Books.
Petzina, D. 1967. *Autarkiepolitik im Dritten Reich. Der national sozialistiche Vierjahresplan*. Stuttgart: Deutsche Verlagsanstalt.
Sarti, R. 1971. *Fascism and industrial leadership in Italy 1919–1940*. Berkeley: University of California Press.
Schweitzer, A. 1964. *Big business in the Third Reich*. Bloomington: Indiana University Press.
Sohn-Rethel, A. 1973. Ökonomie und Klassenstruktur des deutschen Faschismus. Aufzeichnungen und Analysen. Frankfurt am Main: Suhrkamp Verlag. Trans. Martin Sohn-Rethel as Economy and Class Structure of German Fascism, London: CSE Books.
Turner Jr., H.A. 1985. *German big business and the rise of Hitler*. Oxford/New York: Oxford University Press.

Fasiani, Mauro (1900–1950)

Massimo Finoia

Keywords

Business cycles; Consumption taxation; Corporate state; Einaudi, L.; Excise tax; Fasiani, M.; Fiscal illusion; Fisher, I.; Fuoco, F.; Labour supply; Mathematical economics; Pareto, V.; Production at constant costs; Public debt; Public finance; Puviani, A.; Stabilization policy; Tax incidence; Tax shifting; Taxation of income; Taxation of saving; Viti de Marco, A. de

JEL Classifications

B31

Fasiani was born in Turin and died in Genoa. Clearly the most important Italian scholar of fiscal theory to emerge in the interwar period (Buchanan 1960, p. 36), he taught public finance in Turin, Sassari, Trieste and, from 1934, in Genoa. His career was rapid and exclusively academic. Despite his untimely death, he left important works on fiscal theory, and also on economic theory, economic policy and the history of economic thought.

Following Pareto's theory of the ruling class and Puviani's idea of fiscal illusion, which he rediscovered, Fasiani asserts that fiscal activity is to be explained on the basis of the nature of the political entity and not in terms of economic calculus or by sacrifice theories or by the ability-to-pay principle (1932a, 1941). As taxation and public expenditure are political phenomena, it is impossible to know the laws of fiscal activity. Fiscal theory can only be built through static models reflecting the different types of political societies. To De Viti de Marco's models of the 'monopolistic' state, where the ruling class governs only in its own interest, and of the 'cooperative' state, where the ruling class governs in the interest of every member of the community, Fasiani adds the model of the 'modern, nationalistic or corporative' state, in which the ruling class governs in the interest of the collectivity, considered as a whole (1941).

He dealt with the duration of the process of tax shifting (1934) and with the characteristics of intermediate positions in the transition from one state of equilibrium to another (1932b); with tax shifting in conditions of constant, increasing and decreasing costs in competition and in monopoly

(1941, App. I and II) and with the effects of an excise tax under conditions of industrial concentration (1942a). He analysed the different elements determining the 'quantity of labour' and proved the impossibility of understanding the effects of taxation on labour supply assuming as variables only working hours and income (1942c). He devoted much research to the problem of the double taxation of saving (1926), confirming the validity of J.S. Mill's thesis in opposition to the theories of Einaudi and of Fisher. Fasiani also wrote important notes on the application of the Paretian indifference curve apparatus to the classical problem of the relative burden of income tax and consumption tax (1930) and on the analysis of the relationship between taxation and risk-taking (1935b).

In order to study the effects of taxation in a state of equilibrium, Fasiani re-examined and criticized some problems of economic theory. Among other things, he reasserted the hypothesis of production at constant costs and redefined the variables of the labour supply. Specifically he dealt with business cycles and stabilization policy, giving a decisive role to monetary policy (1935a, 1937a, 1942b).

His most important work in the history of economic thought is a very long essay on fiscal theory in Italy (1932c). In this work Fasiani critically examined the general theories of public finance formulated in Italy between 1880 and 1930, that is to say the economic theory, the political theory, the sociological theory, and also the theses on the effects of taxation and public debt on tax shifting and tax incidence.

Finally, the essays on fiscal theory in the 18th century (1936) and on Francesco Fuoco (1774–1841), a forerunner of mathematical economics (1937b), are worthy of note.

Selected Works

A full bibliography of Fasiani's works is contained in: *Rivista di Diritto finanziario e Scienza delle Finanze* 9(September 1950): 216–218.

1926. Sulla teoria dell'esenzione del risparmio dall'imposta. *Memorie della Reale Accademia delle Scienze di Torino* 61, offprint.

1930. Di un particolare aspetto delle imposte sul consumo. *La Riforma Sociale* 41-(January–February): 1–20.

1932a. Temi teorici ed 'exponibilia' finanziari. *La Riforma Sociale* 43(July–August): 383–425.

1932b. Velocità delle variazioni della domanda e dell'offerta e punti di equilibrio stabile e instabile. *Atti della Reale Accademia delle Scienze di Torino* 67: 383–425.

1932c. Der gegenwärtige Stand der reine Theorie der Finanzwissenschaft in Italien. *Zeitschrift für Nationalökonomie* 3(3): 651–691; 4(1) (1933): 79–107; 4(3): 357–388.

1934. Materials for a theory of the duration of the process of tax shifting. *Review of Economic Studies* 1(February), 81–101; 2, February 1935, 122–137.

1935a. Fluttuazioni economiche ed economia corporativa. *Annali di Statistica e di Economia* 3: 1–70.

1935b. Imposta e rischio. In AA. VV., *Studi in onore del prof. Salvatore Ortu Carboni.* Roma: Tipografia del Senato.

1936. Precedenti di alcune recenti teorie finanziarie. *Annali di Statistica e di Economia* 4: 195–240.

1937a. Principi generali e politiche della crisi. *Annali di Statistica e di Economia* 12: 25–108.

1937b. Note sui 'Saggi economici' di Francesco Fuoco. *Annali di Statistica e di Economia* 5: 1–131.

1941. *Principi di Scienza delle Finanze,* 2 vols. Turin: Giappichelli; 2nd ed, 1951.

1942a. La translazione dell'imposta in regime di concentrazione industriale. *Studi Economici Finanziari Corporativi* 2(April–September): 200–225.

1942b. Potenziale di lavoro e moneta. *Annali di Statistica e di Economia* 9–10: 65–137.

1942c. Appunti critici sulla teoria degli effetti dell'imposta sull'offerta individuale di lavoro. *Annali di Statistica e di Economia* 9–10: 139–233.

Bibliography

Buchanan, J.M. 1960. La scienza dell finanze: The Italian tradition in fiscal theory. In *Fiscal theory and political economy, selected essays*, ed. J.M. Buchanan. Chapel Hill: University of North Carolina Press.

Cosciani, C. 1950. Mauro Fasiani. *Economia Internazionale* 3: 913–919.

Einaudi, L. 1950. Mauro Fasiani. *Rivista di Diritto finanziario e Scienza delle Finanze* 9: 199–201.

Scotto, A. 1950. Gli scritti di Mauro Fasiani. *Rivista di Diritto finanziario e Scienza delle Finanze* 9: 202–215.

Faustmann, Martin (1822–1876)

Anthony Scott

Keywords

Faustmann, M.; Forests; Ohlin, B. G.; Rotation periods

JEL Classifications

B31

Faustmann was a German forester who spent much of his life working on the grand-ducal forests of Hesse. Between 1849 and 1865 he entered into controversies with other foresters concerning methods of forest valuations, his ideas eventually prevailing among that minority of forest economists who accepted the discipline of a positive rate of interest in making forest calculations. Although it has been said that his work was approved by such 'national economists' as Wagner and Roscher (*Allgemeine Deutsche Biographie*, 1877), it was evidently quite unknown to the more theoretically oriented German and Austrian specialists in capital and interest. Incorrect solutions to the optimum forest rotation problem were subsequently offered by such economists as Jevons, J.B. Clark and Irving Fisher, in the course of simplified expositions of the idea of the production period of a single investment. Not until the 1950s did economists working outside forestry realize that Faustmann's approach as explained to generations of resistant forestry school students contained a correct approach to the forestry question.

The economists' discovery was sparked by F. and V. Lutz, M. Gaffney, P.H. Pearse and, a few years later, Paul Samuelson. (The literature suggests that some Scandinavian and German economists, notably Ohlin, either knew of Faustmann's formula or worked it out for themselves.)

Faustmann's formula is derived from his investigations into forest values, needed at that time to guide the allocation of landowners' acres between trees and agriculture. His predecessors had consequently attempted to value the soil and the forest separately. In this they failed, partly because they confused stocks and flows. Faustmann cleared this up in 1849 by providing a single forward-looking approach for the present value of the next and future forest crops. As his professional readership required, his formulation also made it possible to take account of expected planting, husbanding, thinning and harvesting net costs during the life of each subsequent stand. He was able to solve his predecessors' problem by showing that the soil value (with which agricultural values are to be compared) is the value of the forest enterprise when it is still bare land, before a crop rotation has been commenced.

Faustmann is known today by resource economists for two by-products of his original perception. First, he showed correctly how to calculate the rotation age that is optimal for the owner in the presence of all expected costs and expected subsequent harvests. Second, by including the expected net discounted returns from subsequent rotations in his value and rotation-age formulae, he took the step that later eluded 20th-century economists, such as Fisher. He included the implicit forgone rent or shadow price of the land. He showed that the effect of doing this is that a given growth-and-harvest cycle will be shorter than economists' analyses would have predicted. Shorter rotations advance the date on which the next and all subsequent rotations will be harvested, thus reducing the effect of waiting on calculated soil values.

Faustmann made subsequent contributions to professional forestry, but they are of little interest today.

Selected Works

1849. Berechnung des Werthes, welchen Waldboden sowie noch nicht haubare Holzbestände für die Waldwirtschaft besitzen. *Allgemeine Forst und Jagd-Zeitung* 25: 441–455. Trans. W. Linnard and included in Gane (1968). Samuelson (1976) contains an extended bibliography.

1877. Faustmann, Martin. In *Allgemeine Deutsche Biographie,* vol. 6. Leipzig: Duncker & Humblot.

Bibliography

Bently, W.R., and D. Teeguarden. 1965. Financial maturity: A theoretical review. *Forest Science* 2: 76–87.

Dickson, H. 1953. Forest rotation. *Weltwirtschaftliches Archiv* 70.

Dowdle, B. (ed.). 1974. *The economics of sustained yield forestry.* Seattle: College of Forestry, University of Washington.

Fernow, B.E. 1902. *Economics of forestry.* New York: T.Y. Crowell.

Fernow, B.E. 1911. *History of forestry.* Revised ed. Toronto: University of Toronto Press.

Gaffney, M. 1960. *Concepts of financial maturity of timber and other assets*, Agricultural economics information series no. 62. Raleigh: Department of Agricultural Economics, North Carolina State College.

Gane, M. (ed.). 1968. *Martin Faustmann and the evolution of discounted cash flow*: *Two articles from the original German of 1849.* Trans. W. Linnard (includes translated articles by von Gehren and Faustmann). Oxford: Commonwealth Forestry Institute; Institute Paper No. 42.

Hiley, W.E. 1930. *The economics of forestry.* Oxford: Clarendon Press.

Lutz, F., and V. Lutz. 1951. *Theory of investment of the firm*. Princeton: Princeton University Press.

Pearse, P.H. 1967. The optimal forest rotation. *Forestry Chronicle* 43: 178–195.

Samuelson, P.A. 1976. Economics of forestry in an evolving society. *Economic Inquiry* 14: 466–492. This article evolved from the symposium edited by B. Dowdle (1974).

Scott, A. 1983. *Natural resources and the economics of conservation*. 3rd ed. Ottawa: Carleton University Press. (See Appendix to ch. 3.)

Fawcett, Henry (1833–1884)

Phyllis Deane

Born on 26 August 1833, the son of a Salisbury draper, Henry Fawcett died on 6 November 1884, by which time he had been Professor of Political Economy in the University of Cambridge since 1863, a Liberal MP since 1864, and Postmaster General under Gladstone since 1880. His political career fulfilled a youthful ambition; his commitment to economics was a consequence of a shooting accident which blinded him at the age of 25. For although he was elected a Fellow of Trinity Hall soon after completing the Cambridge Mathematical Tripos in 1856, the loss of his sight forced him to abandon his studies for the Bar in favour of a professional career which could more easily dovetail with his political preoccupations. He had already begun to read himself into his parliamentary role with the aid of J.S. Mill's *Principles of Political Economy* (1848), and henceforth he depended exclusively on that text to supply the analytical and theoretical framework for his economics.

Fawcett's own textbook, *A Manual of Political Economy* (1863), expounded orthodox classical political economy in the tradition of Adam Smith as updated by Mill. Designed to provide the student (whether undergraduate, politician or general reader) with a clear, relevant, uncomplicated introduction to the state of economic knowledge, and to illustrate its applicability to a changing and complex real world, it went through six diligently revised editions in his lifetime; and his wife, Millicent Garrett Fawcett, a famous suffragette, saw two further editions through the press, the last in 1907. There was much repetition between this work and his other articles and books and the 18 lectures which were his only professorial duty. Fawcett wrote as he spoke, in the spirit of a determinedly non-doctrinaire liberal economist, pragmatically applying the principles of an established discipline to the practical policy problems currently facing government. Prevented by disability from

engaging in systematic research in applied economics, he lacked the interest in abstract reasoning that might have drawn him to theoretical research, where his blindness would have been less of a handicap. Nevertheless, although he chose for himself the role of a teacher, a popularizer of classical orthodoxy, he was intelligently alive to the need to take other considerations into account when prescribing practical policies. For example, his best-seller on *Free Trade and Protection* (1878), after listing all the classical arguments in favour of free trade, went on to defend an Indian five per cent tariff on cotton imports from the United Kingdom, partly on revenue grounds and partly on grounds of natural justice.

The intellectual ferment associated with marginal revolution passed Fawcett by. Yet he did contribute to the debates of the 1860s on the labour question. Mill, for example, took into his *Principles* (with handsome acknowledgement to Fawcett) the idea that unionization was altering behaviour in the labour market by making employers and workers negotiate more rationally. But Fawcett refused to follow Mill in the latter's 1869 recantation of the wages-fund doctrine and took no interest in the 'new political economy' which was exciting the younger generation of Cambridge economists in the late 1870s and early 1880s and on which his successor Alfred Marshall was to set a distinctive personal stamp. On the other hand, his direct, realistic, unpolished attempts to explain the substance and policy implications of elementary economic analysis to non-professionals reached a much wider contemporary audience than the writings of any other late 19th-century English professor of political economy.

Selected Works

1863. *Manual of political economy*. Cambridge.
1865. *The economic position of the British Labourer*. London.
1871. *Pauperism: Its causes and remedies*. London.
1878. *Free trade and protection. An inquiry into the causes which have retarded the general adoption of free trade*. London.

Fawcett, Millicent Garrett (1847–1929)

Murray Milgate and Alastair Levy

A leading suffragist, Millicent Garrett Fawcett was also the author of a widely used elementary textbook, *Political Economy for Beginners* (1870). She married Henry Fawcett in 1867, when he was already Professor of Political Economy at Cambridge, the Member of Parliament for Brighton, and sightless (the result of a stray shot from his father's hunting gun in 1858). This led her to settle down as her husband's full-time secretary. It also brought her at the early age of twenty into close contact with a progressive intellectual circle which included among its elder statesmen Grote and Mill, and also Maurice, Sidgwick and Cairnes. Her first published article, in *Macmillan's Magazine* on Sidgwick's lectures at Cambridge to the unrecognized women students of the day (who included Mary Paley), led to a commission from Alexander Macmillan to write a primer on political economy based on her husband's *Manual of Political Economy*. While her *Political Economy for Beginners* is unremarkable in most respects, it does not follow Mill into the quick-sand of the wages-fund doctrine (see, for example, 1870, p. 25), and it was influential in accelerating that process of establishing economics as a suitable discipline for textbook writers which had been set in motion by Jane Marcet.

Nearly a quarter of a century later, she followed it with *Tales in Political Economy* (1894) which she confessed was little more than a 'plagiarism of Harriet Martineau's idea of hiding the powder of political economy in the raspberry jam of a story' (p. v). The book comprises four stories set on a desert island (thereby inculcating the view that the discipline deals in universals, which some see as having had unfortunate consequences in subsequent years), to illustrate the doctrines of free trade and division of labour, the theory of competition, and the

theory of money. In the latter, coconuts serve as money, and the usual rules of the quantity theory are thereby elucidated in what is, for that theory, a rich institutional setting.

In 1872 she contributed eight of the fourteen chapters to *Essays and Lectures*, a book co-authored with Henry Fawcett. Amongst other topics, she attacked the expansion of the national debt, and opposed the extension of free elementary education on the grounds that it might remove checks to population. In two other essays she promoted the cause of higher education for women, a programme to which she helped to give more concrete form when she was later instrumental in the setting-up of Newnham Hall, Cambridge, which was incorporated as the first women's college in that city's university in 1874.

It was, however, in the area of the struggle for women's citizenship that she played her most significant role. She had joined a suffragist group as early as 1867, but it was only after Henry Fawcett's death in 1884 that she was able to allocate more time to her own political activities. From 1897 until 1918 (the year in which the suffrage was first extended to women in Britain), she was President of the National Union of Women's Societies and after her retirement she continued to campaign for full suffrage (achieved in 1928) and for professional and legal rights. She gave the movement her practical and intellectual support for better than 50 years, a measure of her dedication to the cause.

Selected Works

1868. The education of women of the middle and upper classes. *Macmillan's Magazine* 17(102): 511–517.
1870. *Political economy for beginners*. London: Macmillan.
1872. (With H. Fawcett.) *Essays and lectures on social and political subjects*. London: Macmillan.
1874. *Tales in political economy*. London: Macmillan.
1924. *What i remember*. London: T. Fisher Unwin.

Fay, Charles Ryle (1884–1961)

Murray Milgate and Alastair Levy

Lancashire-born economic historian, whose grandfather worked as a boy on the construction of the first railway coaches for the Liverpool and Manchester Railway and later invented the chain brake used for the emergency stopping of trains, Fay subscribed to a vision of the progress of industrial society towards 'happiness and beauty'. Increased specialization and improvements in the division of labour were, for him, essential to progress. Fay was not, however, unaware that the historical record of industrialization had been marred by hardship, poverty and waste. But these effects had not, in his view, been unavoidable. The exploitation of child and female labour, the appalling conditions in Britain's factories and industrial towns in the 19th century, and the recurrence of distress in agricultural communities, all received Fay's strong condemnation. His liberalism had a social conscience about it. He certainly did not number among those apostles of social laissez-faire who, on his own speculation, might well be found on the lowest ledge of Dante's Inferno (1928, p. 358).

Fay's academic career is easily summarized. He was a favourite pupil of Marshall at Cambridge, and in 1908 he was elected to a fellowship at Christ's College. The same year saw the publication of his study of co-operation in agriculture which established his credentials as an economic historian. Fay remained in Cambridge until 1921, when he removed to Canada to take up a chair in Economic History at Toronto. Nine years later, he returned to Cambridge as Reader in Economic History, where he remained until his retirement.

Some idea of Fay's humane and liberal instincts can be gained from his *Co-operation at Home and Abroad* (1908). Its central thesis was that, contrary to popular opinion at the time, there remained both a social and economic role to be filled by small cultivating ownership. Its prospects, however, rested on the ability of its participants to establish what would today be called marketing boards. Fay saw in

the Canadian wheat pools and the cases of co-operation among Californian fruit growers the promise of things to come (1928, p. 250). Never losing his faith in the market, he stressed that this kind of co-operation was the antithesis of collective ownership and that, what is more, it was the only form of agricultural co-operation that the historical record suggested might work (1908, pp. 350–52). There is more than a faint echo of John Stuart Mill in this advocacy of producer co-operatives over collectivization.

Fay's *Life and Labour in the Nineteenth Century* (1920) expanded on his concern with social history and was based on his Cambridge lectures; it surveyed the main features and figures of the economic, political and social history of the period, and examined the relationship between them and theoretical discourse in economics. This project was repeated on a rather more grand scale in *Great Britain From Adam Smith to the Present Day*, a book first published in 1928 which went through five editions before Fay's death in 1961. This book embodies all the hallmarks of Fay's approach to the study of history. In particular, it reveals very clearly his attempt to trace to their basis in economic theory the practical and political ideas around which history unfolded. In a similar fashion, the subject of protection came under Fay's scrutiny in *The Corn Laws and Social England* (1932) and *Imperial Economy* (1934).

Selected Works

1908. *Co-operation at home and abroad: A description and analysis*. London: P.S. King.
1920. *Life and labour in the nineteenth century*. Cambridge: Cambridge University Press.
1928. *Great Britain from Adam Smith to the present day*. London: Longmans, Green; 5th ed., 1950.
1932. *The corn laws and social England*. Cambridge: Cambridge University Press.
1934. *Imperial economy and its place in the formation of economic doctrine 1600–1932*. Oxford: Oxford University Press.
1940. *English economic history, mainly since 1700*. Cambridge: W. Heffer & Sons.

Fecundity

John L. Newman

Fecundity is defined as the ability to reproduce, whereas fertility is actual reproduction. Because differences in both unobserved fecundity and contraceptive behaviour can cause observed variation in fertility, it can be difficult to separate biological from behavioural influences on fertility. This identification problem is more troublesome in studies of individual than in aggregate fertility behaviour. Fertility trends and differentials at the aggregate level must be due primarily to socio-economic factors since even wide variations in levels of health and nutrition have little effect on fecundity. Only in populations experiencing widespread malnutrition or a high prevalence of diseases leading to sterility (as has occurred in parts of Africa) does fecundity appear to be significantly impaired.

The treatments of fecundity in economic and demographic models of individual fertility behaviour will be compared using a framework that focuses on the stochastic process generating births. The single parameter (p) characterizing a waiting time process generating births is specified as the difference between an underlying component (n) that is exogenous to the individual decisions and the choice of contraception (c). Based on perceived costs and benefits, parents choose c (between 0 and n) to affect their probability of a birth.

Demographers who follow in the tradition of Henry (1957) model the reproductive process and the stages through which a women passes throughout her fertile period, but do not model the choice of c. Such a demographic model of a non-contracepting population can be considered a special case of a more general decision-theoretic model if that model permits the optimal choice of c to be zero. The demographic and economic approaches are therefore not to be distinguished by whether they model the decision for c, but by how they implicitly or explicitly model n, the

underlying component that is exogeneous to the couple's decisions. The main distinguishing features are (1) whether *n* is assumed to be a function of fecundity only or also of various socioeconomic variables and (2) whether *n* is represented by a single (or possibly age-dependent) value or takes on distinct values corresponding to different stages throughout the interval between births. If *n* is solely a function of fecundity, then observed correlations of fertility with socioeconomic variables will reflect only those variables' influence on *c*. If not, the observed correlations will reflect a combined influence on *n* and *c*.

Economic models that focus on the price and income variables affecting fertility typically regard *n* as reflecting the level of fecundity, with variations in *n* being uncorrelated both with socioeconomic variables that explain *c* and with *c*'s error term. The level of fecundity can influence the choice of *c*, whether or not couples can perceive their fecundity. Couples who perceive their higher fecundity may try to offset their higher *n* by choosing a higher *c*. If total contraceptive costs depend on the level chosen, the offset may not be complete and couples with higher *n* may have a higher probability of a birth than otherwise identical couples. The contraceptive decisions of those unable to ascertain *n* will also be affected, to the extent that they choose *c* conditional on their current number of children alive and to the extent that, at any given time, higher fecundity couples have more children.

The possible dependence of *c* on *n* does not present difficulties in estimating the determinants of *n*–*c*. However, the determinants may be estimated only after eliminating the effects of both unobserved fecundity, *n*, and the errors in predicting the choice of *c* from the likelihood function used to describe fertility histories. Provided *n* is uncorrelated with the socioeconomic variables, the combined error terms can be treated as a random effect. The random effect can be integrated out of the likelihood function by assuming a parametric distribution, if results do not appear sensitive to the choice of distribution. If the results are sensitive to the distributional assumption, then a nonparametric procedure may be followed.

While the expected number of births can be derived from the estimated probabilities, the usual procedure has been to regress the number of births on the socioeconomic variables that determine *c*. If *n* is uncorrelated with the latter variables, then the error term of a regression on completed family size will also be uncorrelated with them.

Two potential problems arise when the number of births to those at younger ages and with incomplete families is regressed on socioeconomic variables. The distribution of the error term in the regression may then be misspecified since the number of births reflects the outcomes of waiting-time processes. This is not likely to be a serious problem when couples have had sufficient time for their behaviour to compensate for differences in levels of fecundity.

A more serious problem arises if the observations on fertility histories are censored, as would be expected for younger women. Those couples who have chosen a lower probability of a birth are more likely to have the lengths of their births intervals truncated by the observation date. This imparts a bias to the estimated effects of socioeconomic variables on the number of births. It can be corrected by using additional information on the censored lengths of birth intervals to infer the distribution of uncensored intervals. The likelihood function describing fertility histories is amended to incorporate the probability of not observing a birth, which is equal to one minus the cumulative distribution function of the uncensored distribution. How useful the censored observations are in providing information on the uncensored observations is an issue that must be decided on empirical grounds.

In summary, economic models that assume *n* to be uncorrelated with socioeconomic variables will attribute an observed correlation of fertility with such variables to their influence on *c*. An explicit consideration of fecundity will be required in these models if one is interested in the determinants of the probability of a birth and the spacing of births or if one is interested in the determinants of births and must use observations on women who cannot be assumed to have completed their fertility.

Demographic models of birth probabilities implicitly model n as being correlated with socioeconomic variables. Under this interpretation, an observed correlation between fertility and such variables can exist even when c is always equal to zero (i.e. when couples are not trying to control their births).

The implicit dependence of n on socioeconomic variables is apparent in the analyses of natural fertility populations, defined by Henry as those populations that do not practice contraception or induced abortion. A key technique of these analyses (e.g. Leridon 1977) is to decompose natural fertility into its underlying components which are: (1) the age at marriage and duration of marital separation, (2) the waiting time to conception for a susceptible woman (3) the time added to the birth interval by intra-uterine mortality, (4) the duration of postpartum infecundability and (5) the age at onset of permanent sterility. Differences in gestation lengths are inconsequential. This methodology of breaking down fertility outcomes into intervening components has been extended to the case of contracepting populations by considering (6) the use and effectiveness of contraception and (7) induced abortion. By definition, any determinant of fertility must act through one or more of these proximate determinants.

The first five components interact to yield substantial variations across natural fertility populations in expected mean completed family sizes for women who are married at age 20. The mean family sizes range from 5.4 under the marital fertility rates prevailing in villages near Bombay in 1954–55 to 10.9 for the Hutterite population in the USA with marriages between 1921 and 1930 (Leridon 1977). Based on a sensitivity analysis where the natural fertility components are varied separately through their approximate ranges, Bongaarts and Potter (1983) conclude that the largest variations in simulated total fertility rates are due to changes in the age at marriage and in the duration of postpartum infecundability, both of which can be substantially affected by individual decisions.

Thus, one implication of the demographic approaches is that n is determined by a combination of factors (2) through (5). If n is represented by a single value throughout the birth interval when, in fact, distinct biological factors operate over different stages of the birth interval, the model will be misspecified. The possible specification error must be balanced against the bias that would arise if the identification of the different stages is accomplished by conditioning on a choice variable of the parents, such as the length of breastfeeding.

A second implication is that n is a function of socioeconomic variables. If both n and c are functions of the same variables, then a non-contracepting population can be identified solely on the basis of fertility data, only under a maintained hypothesis that parents initiate or alter their control after a birth. This hypothesis is maintained in the literature on natural fertility. Identifying the effects of observed socioeconomic variables requires an explict formulation of how the variables affect n and c, noting that c may also depend on n. Identification may be facilitated if either economic theory, or a biological theory of the determinants of n, specifies how n and c respond to births and deaths.

Treating the unobserved components of n–c as a random effect, as described above, will provide a reduced form estimate of the effect on fertility of a variable that affects both n and c. Identifying the separate effect on c is possible if n can be treated as a fixed effect and eliminated from the estimating equation. A comparison of the estimated coefficients from the fixed effect model and the random effect model would provide information on the variable's effect on n.

See Also

- ▶ Demography
- ▶ Family Planning
- ▶ Fertility

Bibliography

Bongaarts, J., and R.G. Potter. 1983. *Fertility, biology, and behavior: An analysis of the proximate determinants*. New York: Academic Press.

Bulatao, R.A., and R.D. Lee (eds.). 1983. *Determinants of fertility in developing countries*, 2 vols. New York: Academic Press.
Henry, L. 1957. Fécondité et famille: modèles mathématiques (I). *Population* 12(3). Trans. as 'Fertility and family: Mathematical models I'. In *On the measurement of human fertility*, ed. M.C. Sheps and E. Lapierre-Adamyck. New York: Elsevier, 1972.
Leridon, H. 1977. *Human fertility: The basic components*. Chicago: University of Chicago Press.

Federal Reserve System

Donald D. Hester

Abstract

The Federal Reserve System was established in 1913 to provide the United States with an elastic currency. It managed security offerings to finance the First World War, and evolved from a set of 12 semi-autonomous banks to a centralized institution in the 1920s. Having failed to prevent the Depression of the early 1930s, it was substantially reorganized in 1933 and 1935. After the Second World War and a 1951 accord reached with the Treasury, it started on an odyssey of monetary policy interventions, employing many policy instruments, indicators, and powers with varying degrees of success to the present day.

Keywords

'Availability of credit' doctrine; Accord (1951); Availability of credit; Bank holding companies; Bank Holiday; Central banking; Commercial paper; Deposit turnover; Depository Institutions Deregulation and Monetary Control Act; Discount rate; Discount window; Discretionary monetary policy; Eligible paper; exchange rates; Federal funds; Federal reserve system; Federal Open Market Committee (FOMC); Financial deregulation; Financial holding company; Financial Services Modernization Act; Free gold; Garn–St Germain Act; Glass–Steagall Act; Gold standard; Great Depression; Inflation; Meltzer, A.; Money supply measures; Net free reserves; Open-market operations; Real bills doctrine; Repurchase agreements; Reserve requirement; Retail sweep programmes; Riefler–Burgess doctrine; Unemployment

JEL Classifications

B31

The Federal Reserve System of the United States was established on 23 December 1913, when President Woodrow Wilson signed the Federal Reserve Act. The need for a new federal banking institution became clear when a severe crisis occurred in 1907. In May 1908 the Aldrich–Vreeland Act established a bipartisan National Monetary Commission that proposed establishing a National Reserve Association with 15 locally controlled branches that would 'provide an elastic note issue based on gold and commercial paper' (Warburg 1930, p. 59). The proposal was not enacted, nor was a subsequent proposal for a central bank with about 20 branches that would be controlled by a centralized Federal Reserve Board, consisting largely of commercial bankers. In the debate preceding the Federal Reserve Act, banking industry domination was rejected in favour of a board that had five members appointed by the President and two ex officio members, the Secretary of the Treasury and the Comptroller of the Currency. The appointed members had staggered terms and were to represent different commercial, industrial, and geographic constituencies. A sixth appointed member representing agriculture was added in 1923. The composition of the Board and its relation to Federal Reserve banks were drastically changed in 1935. Partly because of continuing disagreements about public versus commercial bank control, the new Board's powers were left ambiguous in the act.

The act mandated that all national banks become members of the new system and stockholders of Federal Reserve banks. Because reserves were to be concentrated in 12 Federal Reserve banks, the act substantially reduced reserve requirements at national banks. State

chartered banks could join if they chose to and were judged to be financially strong. The first Board was sworn in on 10 August 1914 and the system opened for business on 16 November 1914. Federal Reserve notes that were backed 100 per cent by 'eligible paper' and, additionally, 40 per cent by gold began to circulate. Eligible paper was self-liquidating, short-term paper that arose in commerce and industry. The rationalization for eligible paper was the real bills doctrine, which held that credit extended for financing only the production and distribution of goods would not lead to inflation. The doctrine is invalid because of fungibility; there is no relation between paper acquired by Federal Reserve banks and loans the commercial banks are extending. In addition, all deposits at Federal Reserve banks had to be backed at least 35 per cent by gold. Subsequent amendments to the act effectively eliminated the supra-100 per cent collateralization of notes. A June 1917 amendment to the act forced all member banks to pool required reserves at Federal Reserve banks and further reduced reserve requirements to decrease the burden of membership on national banks and attract more statechartered banks to the system.

The Early Years

The early years of the Federal Reserve System were marked by struggles to define the distribution of power between Federal Reserve banks and the Board, in the context of growing US involvement in the First World War. The Board gradually assumed more powers, but was unsuccessful in controlling open-market trading, which inevitably was concentrated in New York. Benjamin Strong, the New York bank governor, managed system trading. (Until 1935 the chief executives of Federal Reserve banks were called 'governors'. After 1935 their title was changed to 'president' and members of the Board were called 'governors'.) The Federal Reserve System was made fiscal agent for the Treasury in 1920, but the Treasury dealt directly with Federal Reserve banks, not the Board. Until 1922 the Board's statistical research office was located in New York, and arguably the Board was less informed than the New York bank about money market conditions.

Federal Reserve banks immediately sought earning assets in order to pay expenses and the six per cent required dividends on member bank capital subscriptions. As they expanded their portfolios of bills, US securities, discounted commercial paper, and acceptances, the breadth and liquidity of these markets increased. In early 1915 the New York bank was buying and selling for other Federal Reserve banks. Discount rates charged by reserve banks varied across Federal Reserve districts.

In anticipation of the US declaration of war on Germany in 1917, Federal Reserve banks became responsible for issuing and redeeming short-term Treasury debt certificates before and during Liberty Loan drives. There would be four large Liberty Loans and a Victory Loan in 1919 that required extensive Federal Reserve involvement. US bonds were sold to the public on an instalment plan by member banks; the interest rate banks charged on the unpaid balance on a bond was equal to the coupon rate on the bond. Member banks, in turn, discounted short-term US debt at Federal Reserve banks at an interest rate below the yield on the debt, which allowed them to recover their costs of instalment lending.

US government interest-bearing debt rose from $1.0 billion at the end of 1916 to $25.5 billion at the end of 1919, and would never again fall below $15 billion. This huge increase, and the fact that Federal Reserve banks offered preferentially low interest rates when member banks discounted government debt, had important lasting consequences on the money market. Before the war, Federal Reserve banks had schedules of discount rates that varied across the quality and maturity of discounted paper and the amount of borrowing by a member bank. Because of the low discount rate on government debt, member banks almost exclusively offered it as collateral when borrowing. The discount rate effectively became the rate charged on government debt. By 1922 each reserve bank effectively had a single discount rate, but rates still varied across Federal Reserve districts.

The November 1918 armistice brought new challenges. Continuing shortages of food and

other goods in Europe and large increases in the stock of money led to inflation in the United States. The rate of inflation peaked in May 1920 and was followed by a sharp deflation in the following year of about 45 per cent in wholesale prices. In that year industrial production fell by about 30 per cent and unemployment soared. Until October 1919 Federal Reserve banks were obliged to keep the low wartime discount rates in order to allow banks and the public to absorb the 1919 Victory Loan. In November, Federal Reserve banks began raising their discount rates in an effort to combat inflation. In June 1920 four banks raised the rate to seven per cent. Amplifying the effects of the interest rate increases was an outflow of gold to Europe and a sharp reduction in discount window borrowing as Federal Reserve banks cut back on subsidizing the public's instalment purchases of US bonds.

The Boston bank lowered its rate from seven per cent to six per cent in April 1921, and was gradually followed by other reserve banks in an effort to respond to the slowdown. Deposits at all member banks reached a local maximum of $26.1 billion in the December 1919 call report and then fell to $22.8 billion in the April 1921 report. Discount window borrowings reached a year end high of $2.7 billion in December 1920 and then fell to $0.6 billion at the end of 1922 as gold flows turned positive. As gold flowed in, reserve banks lowered their discount rates to 4.5 per cent in 1923 and early 1924.

While gold inflows slackened after 1923, it became apparent that new operating guidelines were needed. Governor Strong understood that the real bills doctrine was invalid and that many countries were not acting according to the old gold-standard rules. As interest rates fell, most reserve banks were again acquiring securities to augment their income. Strong, on the other hand, had begun to sterilize the New York bank's holdings of gold by selling its securities in the open market. The Treasury was concerned that reserve bank trading was upsetting securities markets when it was buying or selling debt. In May 1922 the reserve banks established the Governors Executive Committee consisting of the governors of the Boston, Chicago, Cleveland, New York, and Philadelphia banks to manage transactions for all 12 banks. The committee executed orders on behalf of the banks in the light of Treasury plans and made recommendations, but acted only as agents and had no executive power. In April 1923 it was renamed the Open Market Investment Committee (OMIC), which had the same membership as its predecessor but was required

> to come under the general supervision of the Federal Reserve Board; and that it be the duty of this committee to devise and recommend plans for the purchase, sale and distribution of open-market purchases of the Federal Reserve Banks in accordance with. . .principles and such regulations as may from time to time be laid down by the Federal Reserve Board. (Chandler 1958: 227–8)

Strong dominated the OMIC and began to understand the way open-market operations worked. He noted in particular that the sum of reserve bank open-market purchases and gold inflows almost equalled negative changes in member bank borrowing. He developed a case for active monetary policy and argued that restrictive monetary policy should be initiated with open-market sales and followed by increases in the discount rate. This was the likely origin of member bank borrowings and nominal interest rates as indicators of monetary policy. Policy instruments were open-market operations and the discount rate. While proposals to change discount rates originated with Federal Reserve banks, they required Board approval, which may explain why Strong preferred to lead with open-market operations. Strong was sensitive to the effects of monetary policy on prices, but objected to any legislated targeting of prices. His analysis was seriously incomplete when banks were not net borrowers from the Federal Reserve, and in such circumstances so were his policy tactics. Tragically, beginning in 1916 Strong suffered from recurrent attacks of tuberculosis and would die in October 1928, before such circumstances arose.

The 1923 Board Annual Report advocated an activist policy, but continued to support the real bills doctrine. In response to pressure from the Treasury and the Board, Federal Reserve banks

sold most of their government securities in 1923; yearend holdings fell from $436 million to $134 million between 1922 and 1923. Federal Reserve notes and member bank reserves backed by such assets were unjustifiable under the doctrine, and the Treasury objected to Federal Reserve banks profiting from such assets. However, at the end of 1924 the banks held $540 million, and the banks' portfolio of government securities fluctuated considerably in the following years in response to changes in the volume of discounted bills and gold flows. Discount rates at Federal Reserve banks were lowered in the latter half of 1924 and 1925 before converging on four per cent at the beginning of 1926, largely following short-term interest rates in New York. Short-term market rates fell because of a sharp recession; the Federal Reserve index of industrial production (1997 = 100) fell from 7.84 in May 1923 to 6.43 in July 1924. Clearly policy was active, but not because of the real bills doctrine!

The discount rate was four per cent in June, when Federal Reserve banks began to cut the rate to 3.5 per cent and to make open-market purchases. At the beginning of 1928 discount rates were increased because of developing speculation in the stock market and continued to rise to as much as six per cent in October 1929, when the stock market crashed. In part, Federal Reserve discount rates were again responding to changes in industrial production, which had been quite sluggish until the end of 1927 and then began to grow rapidly until July 1929. In part, the 1927 rate cut reflected Federal Reserve efforts to help the United Kingdom maintain sales of gold at the pre-war sterling price, which had been restored in 1925. Governor Strong and Montagu Norman, the Governor of the Bank of England, were working to reestablish a gold standard that could restore order to international finance. To help the United Kingdom in 1925, the New York bank extended the Bank of England a $200 million gold credit and attempted to keep interest rates low in New York relative to those in London. By reopening gold sales at the pre-war price, Britain had effectively revalued the pound upward in 1925 by about ten per cent, with devastating consequences for its economy.

As Strong's health failed in 1928, a leadership vacuum developed. In an attempt to coordinate policy among all 12 reserve banks and the Board, the Board proposed in August 1928 that the five member OMIC be replaced by a new Open Market Policy Committee (OMPC) that included all 12 reserve bank governors and was chaired by the Governor of the Federal Reserve Board. This proposal was rejected by bank governors, but a modified form was adopted in January 1930. Strong had been aware of growing stock market speculation and did not object to Federal Reserve open-market sales and the increase in the discount rate. These actions were reinforced by outflows of gold. In mid-1928 gold flows reversed, apparently attracted by high and rising short-term interest rates. Federal Reserve banks continued to sell bills and government debt, forcing member banks into the discount window to the extent of about $1 billion in the second half of 1928 and in the middle of 1929. At the end, Strong was aware of the danger of restrictive monetary policy actions over an extended period on the real economy, but remained reasonably optimistic that the situation could be controlled (Chandler 1958: 460–3). After his death the struggle for control continued between his successor at the New York bank, George L. Harrison, and the Board; the latter argued that the real bills doctrine was not dead and that reserve banks should take direct action to penalize member banks making loans that supported security speculation. The Federal Reserve index of industrial production peaked in July 1929, Bureau of Labor Statistics (BLS) wholesale and consumer price indices had been slowly falling since 1926, and in October the stock market collapsed.

The Great Depression

Led by the New York bank, the Federal Reserve flooded the money market with cash by aggressively buying government securities. Discount window borrowing by member banks fell from

$1037 million in June 1929 to $632 million in December and to $271 million in June 1930. Further, discount rates at reserve banks were rapidly reduced; at the New York bank the rate was lowered from six per cent in October to 2.5 per cent in June 1930. The monthly average Standard and Poor common stock index (1935–1939 = 100) began to stabilize; it was 195.6 in January 1929, 237.8 in September, 159.6 in November, and 191.1 in April 1930. However, the index of industrial production continued to fall after the open-market purchases, and the BLS index of wholesale prices was ten per cent lower in 1930 than in 1929.

In mid-1930 reserve banks sharply reduced their purchases of government securities in the belief that monetary policy was adequately expansionary. The OMPC seems to have been guided by what Meltzer (2003: 164) calls the Riefler–Burgess Doctrine: 'If [discount window] borrowing and interest rates were low, policy was easy; if the two were high policy was tight.' An interpretation is that if member banks wanted to lend they could have inexpensive and relatively easy access to funds; if not, there was little more that the Federal Reserve could do. While total member bank discount window borrowing was positive, many banks were holding excess reserves. Conventional wisdom has it that the reserve banks should have continued buying securities. However, it is unclear even today whether continued large open-market purchases by the Federal Reserve would have had much of an impact on real economic activity in late 1930; the experiment was never tried. Rapid expansion of reserves and member bank deposits did occur in the late 1930s, with little effect on real economic activity.

On average about 600 bank failures a year occurred between 1920 and 1930; most failing banks were small and not members of the Federal Reserve System. The number of failing banks doubled in 1930 and increased by another 70 per cent in 1931. The total deposits of failing banks between 1920 and 1930 averaged less than $200 million a year, but more than quadrupled in 1930 and doubled again in 1931. Total deposits and currency had begun to fall after December 1928 and continued to fall after the stock market crash. Currency in circulation began to rise in November 1930, as bank failures increased. Industrial production and wholesale prices were falling at an accelerating rate. The directors of the New York bank counselled Governor Harrison to continue open-market purchases in 1930, but he encountered opposition in the OMPC and little was done. Net gold inflows were offset by open-market sales because the OMPC collectively believed monetary policy was expansionary. Reserve bank discount rates and money market interest rates trended down until 21 September 1931, when the United Kingdom suspended gold payments.

The British abandonment of gold led to very large withdrawals of gold and currency from the United States that were initially partially offset by open-market purchases of bills and increased discount window borrowing, which occurred at sharply higher interest rates as recommended by Bagehot (1873). However, Federal Reserve bank credit fell from $2.2 billion in October 1931 to $1.6 billion in March 1932. During this period of rising bank failures, rapidly declining economic activity, and falling prices, Harrison argued against open-market purchases for a number of reasons, but primarily because of the possibility of a shortage of 'free gold', that is, gold that was not required as collateral for Federal Reserve notes and reserves. The Glass–Steagall Act of 1932 authorized the Federal Reserve banks temporarily to use US government securities as collateral for Federal Reserve notes and thus largely solved the problem of a lack of free gold. In February 1932 Federal Reserve banks began aggressive open-market purchases of government securities that more than offset continuing gold losses and allowed member bank borrowings to fall about 50 per cent by August 1932. Discount rates at the New York and Chicago banks were lowered to 2.5 per cent in June 1932, but all other banks kept their rates at 3.5 per cent until the national banking 'holiday' that began on 5 March 1933 when President Roosevelt closed all US banks. Net free reserves (excess reserves minus discount window borrowing) had turned positive in September and thus signalled excessive ease to some individuals on the OMPC.

Restructuring the Federal Reserve System

It was obvious that the Federal Reserve had been ineffective in combating the collapse of the banking system and responding to the Great Depression. The banking system and the Federal Reserve needed to be restructured and strengthened. The Emergency Banking Act of 9 March 1933 authorized the Treasury to license and reopen national banks that were judged to be sound; state chartered banks that were sound would receive licences from state banking commissioners. Many reopening banks received capital injections by selling preferred stock to the Reconstruction Finance Corporation. At year end 1929 there were 24,026 commercial banks of which 8522 were members of the Federal Reserve System; at year end 1933 there were 14,440 commercial banks of which 6011 were member banks. For a period of one year all banks, whether members or not, could borrow on acceptable collateral from Federal Reserve banks.

Many of the reforms that were adopted would survive at least until late in the 20th century. Because of a belief that the collapse lay in undisciplined stock market trading, the Glass–Steagall Act of 1933 required that commercial banks divest themselves of investment banking activities. This act introduced deposit insurance that became effective in January 1934. It also banned interest payments on demand deposits and allowed the Board to impose ceilings on interest rates that banks could pay on time and savings deposits. Finally, the act renamed the OMPC the 'Federal Open Market Committee' (FOMC), but as in earlier incarnations its executive committee remained the same. The Securities Exchange Act of 1934 authorized the Board to impose margin requirements on stock market trades. Federal Reserve banks were authorized to make commercial and industrial loans to non-financial firms.

Having failed to expand reserve bank credit between July 1932 and February 1933, the Board found itself under extraordinary political pressure to expand resources to the banking system. As Meltzer (2003: 435–41) explains, President Roosevelt threatened to have the Treasury issue currency in the form of greenbacks if the FOMC failed to expand sufficiently. Net free reserves turned positive in May 1933 and rose to more than $3.0 billion by January 1936. The revaluation of gold in February 1934 together with subsequent large gold inflows from Europe and hesitancy to lend by member banks contributed to this surge in excess reserves.

The reconstruction of the Federal Reserve System continued with Roosevelt's nomination of Marriner Eccles to become Governor of the Federal Reserve Board in November 1934. Eccles had argued that system power should be concentrated in the Board and that reserve banks be prevented from undertaking open-market operations on their own accounts. Eccles's initiatives were opposed by Senator Carter Glass, many reserve bank governors, and the banking industry, but he largely succeeded in achieving his goals. The reforms were in the Banking Act of 1935, which restructured the Board to consist of seven appointed governors, each with a staggered 14-year term. The FOMC was restructured to consist of the seven governors and five reserve bank presidents. Two of the governors were to be appointed for four year terms as chairman and vice-chairman of the Board by the president, with the advice and consent of the Senate. Eligible paper was no longer restricted to being short-term paper that originated in commerce and industry. The Board was empowered to vary reserve requirements; the upper limit was twice the percentages that were specified in the 1917 amendments to the Federal Reserve Act.

Members of the renamed Board of Governors of the Federal Reserve System took office in February 1936, with Eccles as chairman. For some time the FOMC had expressed concern about the inflationary potential of large excess reserves. In particular, because excess reserves exceeded reserve bank credit, the FOMC would not be able to absorb them without an increase in reserve requirements. Employing its new policy instrument, on 14 July 1936 the Board announced an increase in reserve requirements on August 15 of 50 per cent on all deposits at member banks. The increase was expected to absorb less than half of

system excess reserves and was not expected to impinge on member bank lending or the economic recovery. In part because of continuing gold inflows, excess reserves were $3.0 billion at the end of July 1936, and averaged about $2.0 billion through the end of February 1937. Because excess reserves continued to be large, the Treasury began to sterilize gold inflows in December 1936, but not to the extent desired by the Board. At the end of January the Board announced a further two-step increase in reserve requirements of one-third to take place in March and May 1937. These actions took reserve requirements to their legal maxima and reduced excess reserves to below $800 million in summer months. In August and September reserve banks reduced their discount rates to one per cent or 1.5 per cent, levels that would last until December 1941. Coinciding with the May increase, the industrial production index.

(1997 = 100) reached a high of 10.4 and then decreased to 7.0 in May 1938. Continuing gold inflows and the Treasury's February 1938 abandonment of gold sterilization allowed excess reserves to increase to $1.5 billion in March 1938. Beginning after the Board's reduction in reserve requirements of more than ten per cent in April 1938, excess reserves began a rise to nearly $7 billion in late 1940; however, industrial production did not pass its 1937 peak until October 1939, after the Second World War had begun in Europe.

Second World War and Recovery

As the war approached gold flowed into the United States, and the FOMC allowed its security holdings to fall and their maturity to lengthen. In response to inflationary pressures, the Board introduced consumer credit controls in September 1941 and again raised reserve requirements to their legal maxima in November. After the United States declared war, monetary policy was constrained to facilitate war finance. In April 1942 the FOMC set interest rate ceilings on treasury bills at 0.375 per cent and on long-term bonds at 2.5 per cent. The yield curve was upward-sloping and effectively 'pegged' by these two boundary conditions into the post-war period. Because capital gains could be earned by buying high coupon securities and selling as they approached maturity, the cost of intermediate term debt was higher than rates shown on the yield curve. Discount rates were lowered to one per cent by all reserve banks and were not raised again until 1948. A preferential discount rate of 0.5 per cent was charged for loans collateralized by short-term US debt. Reserve requirements for central reserve city member banks were lowered in 1942, causing interest-free reserves to disappear into interest-bearing US securities. Finally, a variety of selective credit controls were imposed during and after the war, which ended in August 1945.

Yearend deposits and government securities of member banks had risen from $61.7 billion and $19.5 billion in 1941 to $129.7 billion and $78.3 billion respectively in 1945. Because of the pegging of the yield curve, Federal Reserve bank yearend ownership of US securities rose from $2.3 billion in 1941 to $24.3 billion in 1945; treasury bills were $10 million in 1941 and $14.4 billion in 1946.

The preferential discount rate was eliminated in the spring of 1946. In July 1947 the FOMC relaxed the rate ceiling on treasury bills and the rate rose to about one per cent by yearend. Reserve banks raised the discount rate to 1.25 per cent in early 1948. Eccles's long term as chairman ended in February 1948, but he continued as a member of the Board. Reserve requirements were increased in 1948 as the Board sought to control inflation, although prices were actually falling at yearend when a recession occurred. Indeed, the reserve requirement policy instrument was used many times between April 1948 and February 1951 because it was perceived not to have a direct effect on treasury interest rates. A continuing struggle between the Board and the Treasury for an independent monetary policy would not be resolved until a spurt of inflation after the start of the Korean War led to an accord signed on 4 March 1951. It effectively freed the Board from pegging interest rates. Partly because of frictions leading to the accord, a new chairman, William McChesney Martin, Jr., was appointed in April.

Resumption of Discretionary Monetary Policy

In the Martin era of discretionary monetary policy, new operating techniques were needed. In 1953 the FOMC settled on a policy of 'bills only', which meant that open-market operations would be largely confined to the market for treasury bills, because it was recognized that large policy actions in thin markets could impair market efficiency. Indicators of monetary policy continued to be net free reserves and market interest rates. Because evidence was lacking that interest rates had much effect on private sector investment, a new paradigm, the 'availability of credit' doctrine, was used to rationalize the transmission of policy actions to the real economy. It argued that banks rationed credit to marginal borrowers when restrictive policy led to rising interest rates or indebtedness at the discount window. With these adjustments the FOMC vigorously and unsuccessfully pursued goals of lowering inflation and combating unemployment in the turbulent decade of the 1950s. In that decade there were three business cycles, which were marked by successively rising peaks of interest rates, inflation, and unemployment. The reason for this failure was thought to be inflation-induced rising marginal rates of taxation, which were addressed by large tax cuts in the following decade.

As interest rates rose, the opportunity cost of holding excess reserves rose, which led to the reappearance of a federal funds market in which banks traded reserves. Because banks paid no interest on demand deposits, there was also rapid expansion of the market for commercial paper in which large firms with good credit ratings traded idle funds without the direct intervention of banks. Both markets had atrophied after the 1920s because of low interest rates, and served to change the relation between open-market operations and real economic activity. They were precursors of a wave of innovations that would have similar effects in the coming decade. These included large-denomination negotiable certificates of deposit, one-bank holding companies, offshore 'shell' branches, the Eurodollar market, and bankrelated commercial paper.

Beginning in 1961, the Kennedy administration attempted to coordinate fiscal and monetary policy by proposing large tax cuts to encourage investment and economic expansion. A new problem was that the United States was experiencing large gold outflows as the world continued to recover from the world war. To cope with this new approach and problem, the FOMC was encouraged to abandon its bills-only policy and to attempt to twist the yield curve by buying long-term bonds and selling bills. As short-term rates rose the Board repeatedly raised the ceiling on interest rates that banks could pay on time and savings deposits. It was argued that lower long-term interest rates would encourage capital formation and that higher short rates would discourage foreign interests from converting dollars into gold, as they were entitled to under the Bretton Woods agreements. These efforts were not successful in discouraging gold outflows, but investment and the economy expanded strongly. In 1965 the Board introduced a Voluntary Foreign Credit Restraint programme, which discouraged banks from overseas lending that was not financing US exports. Nevertheless, gold continued to flow out and the requirement that Federal Reserve notes and reserves be backed by gold was cancelled in 1968. Large open-market purchases had been needed to offset gold losses.

Policy coordination between the Board and the new Johnson administration effectively ended in December 1965, when the Board approved an increase in the discount rate because of inflation arising from mobilizing for the Vietnamese War. Net free reserves had turned negative in 1965 and were increasingly so until late 1966. Short-term interest rates rose until October. Higher rates increased the cost of the mobilization and had devastating effects on residential construction and the savings and loan associations and mutual savings banks (hereafter thrifts) that financed it, because in September Congress passed legislation limiting interest rates that thrifts could pay on time and savings accounts. These limits meant thrifts would experience withdrawals of funds or 'disintermediation' because depositors switched funds to government securities, which had no limits. This policy transmission channel would soon

disappear because Congress and the administration could not withstand the resulting political pressures. In 1968 the Federal National Mortgage Association was privatized and in 1970 the Federal Home Loan Mortgage Corporation was created. Both bypassed depository institutions by securitizing mortgage loans. Banks also responded to Board policies and restrictions on innovations by opening overseas offices that were not subject to them. A ten per cent income tax surcharge in 1967 was insufficient to stop inflation, and short-term interest rates rose to new highs in January 1970, when Chairman Martin's term ended. Net free reserves averaged about a negative $1 billion between May 1969 and July 1970. A decrease in short-term interest rates followed the then largest-ever US bankruptcy of the Penn Central Transportation Company in June 1970, but led to large new capital outflows in 1971 that pressured the dollar. The FOMC responded by forcing short-term rates and net borrowed reserves up again.

Towards Flexible Exchange Rates

The amplitude of changes in interest rates increased between 1965 and 1971, and the United States experienced a recession in 1970. As in the 1950s the Federal Reserve was unable simultaneously to achieve satisfactory unemployment, inflation, and exchange rate outcomes. Many of the Board's policy instruments, such as the discount rate, reserve requirement changes, and many regulations had effectively been disabled by innovations, so that only open-market operations were available to achieve multiple targets. For example, an increase in reserve requirements induced banks to resign from the system or to conduct more of their business overseas. One exception to this loss of powers was the 1970 amendments to the Bank Holding Company Act, which finally gave the Board regulatory authority over one-bank holding companies. In August 1971 the Nixon administration, with new Board Chairman Arthur F. Burns as an advisor, announced a 90-day freeze on prices and wages, suspension of gold sales, and several other major changes in the United States. The suspension of gold sales led to a floating exchange rate system, devaluation of the dollar, and sharp rises in dollar-denominated prices in international markets. The shift from a fixed to a floating exchange rate system is likely to have increased the potency of monetary policy, as was predicted by Mundell (1961). The FOMC responded to consequent high inflation by driving nominal short-term interest rates to very high levels in 1973 and 1974, which helped to induce a severe recession beginning in August 1973, but were inadequate because on average the real federal funds interest rate (calculated with the GDP deflator) was negative between the end of 1973 and 1978. Real estate and other durable goods prices rose relative to the GDP deflator, and the international value of the dollar fell. After the resignation of President Nixon in 1974, Congress required the Chairman to explain policy in semi-annual public hearings and report the FOMC's targets for two money stock measures: M1, a measure of transactions balances, and M2, a measure of liquid assets. Friedman and Schwartz (1963) had recommended using money as an indicator of monetary policy instead of interest rates or net free reserves.

Part of the explanation for the policy failure was continuing financial market innovation. Foreign banks operating in the United States grew rapidly and were unregulated until the 1978 International Banking Act, which placed them under Board supervision. The introductions of money market mutual funds (MMMFs) and negotiable order of withdrawal (NOW) accounts in 1972, the Chicago Board Options Exchange in 1973, and financial futures markets in 1975 again began changing the relation between financial and real markets. A more important change was the rapid expansion of repurchase agreements after 1970. In a repurchase agreement, a client's deposits are borrowed to finance a bank's or dealer's inventory of government securities, often only overnight. Large bank holdings of government securities often represented transactions balances of large corporations and state governments that could not easily be controlled.

The real federal funds rate turned distinctly positive in the third quarter of 1979 when Paul

A. Volcker became chairman. In early October he announced that the FOMC would no longer limit fluctuations in short-term interest rates and would use open-market operations to control bank reserves. This was a major policy change from practices dating from the 1951 accord. Further, he imposed eight per cent marginal reserve requirements on non-deposit liabilities, that is, Eurodollar borrowing, federal funds purchased from non-member banks, and funds acquired through repurchase agreements. These vigorous actions together with large income tax cuts by the Reagan administration between 1981 and 1983 drove real short-term interest rates to levels not seen since the early 1930s and caused MMMFs to grow rapidly. In only two quarters between 1979 and 1986 was the average real federal funds less than five per cent. These high rates caused the trade-weighted value of the US dollar to appreciate by 87 per cent between July 1980 and February 1985, which savaged US exports and attracted imports with adverse consequences for US manufacturing.

Financial Deregulation

The landmark Depository Institutions Deregulation and Monetary Control Act was signed by President Carter at the end of March 1980. It radically changed the Federal Reserve System by eliminating the significance of membership in the system. After an 8 year phase-in period, all depository institutions would be subject to uniform reserve requirements on demand and time deposits, although the requirement on the first $25 million of transactions deposits was less than that on other transactions deposits. The Board could vary reserve requirements. All depository institutions had access to reserve bank discount windows. This strengthened the system because banks could no longer threaten to leave it in order to get the lower requirements that many states imposed. Further, Federal Reserve banks were required to charge banks for the cost of services they provided. Before this act they had been giving away services as an inducement for banks to stay in the system. This pricing requirement in turn forced depository institutions to begin to charge their clients for services, which changed the way banking services were used. The act mandated that interest rate ceilings on time and savings accounts be eliminated after six years, increased deposit insurance, and had other important provisions that are beyond the scope of this discussion.

In late 1980 the Board announced that transfers from overseas branches to the United States could be treated as collected funds on the day they were transferred. Before then, transfers in a day were not 'good funds' until the following day. The expansionary effects of this change, rapidly growing repurchase agreements, and other innovations are evident in demand deposit turnover statistics that the Board reported from 1919 until August 1996. Turnover is the annualized value of all withdrawals from deposit accounts divided by aggregate deposit balances.

High interest rates were savaging thrift institutions, which had negative gaps (more fixed-rate assets than fixed-rate liabilities on most future dates), and allowed MMMFs to expand rapidly. Congress intervened in September 1982 by passing the Garn–St Germain Act, which provided temporary emergency assistance and among other changes introduced money market deposit accounts and super NOW accounts, which paid market interest rates. MMMF growth was slowed by this act, but the weakening condition of banks and thrift institutions would result in large numbers of failures as the decade wore on. Large banks also experienced large losses because the appreciating dollar had resulted in failures of sovereign states, especially in Latin America, to meet their loan obligations. Chairman Volcker was heavily involved in negotiating solutions for these defaults.

The restrictive monetary policy resulted in the deepest recession since the Depression; the unemployment rate was 10.8 per cent at the end of 1982. At the end of Volcker's term in August 1987 the unemployment rate had fallen to six per cent and the consumer inflation rate was less than two per cent. Real interest rates had fallen from 10.5 per cent in mid-1981 to four per cent, and the trade-weighted value of the dollar fell correspondingly. Volcker's February 1987 statement of monetary policy objectives to the Congress reported

that M1 was not a reliable indicator of monetary policy and would be de-emphasized.

While his successor, Alan Greenspan, inherited a much improved economy, many problems remained from a rising wave of bank failures and the collapse of thrift institutions. Real estate markets were especially disorderly when the thrift crisis was resolved beginning in 1989 and were further distorted by provisions in the Tax Reform Act of 1986, which disallowed many interest tax deductions. After 1990 interest on home loans was effectively the only deductible interest on individual income tax returns. In addition, a collapse of stock prices in October 1987, strong foreign demand for US currency associated with the collapse of the Soviet Union, and a recession at the end of 1990 presented further challenges. The FOMC responded to these challenges by varying the real federal funds rate, defined using the contemporaneous GDP price deflator inflation rate. This rate fell sharply for two quarters after the stock market crash, rose before falling for two quarters after a second stock market dip in October 1989, and then began to fall in the fourth quarter of 1990. In July 1993 testimony before Congress, Greenspan disclosed that the FOMC was downgrading M2 as an indicator of monetary policy and, as could have been surmised from its actions, that an important guidepost was now real interest rates. The real federal funds rate averaged less than one per cent in 1993. In early 1995 it had risen to four per cent and held that value as an average until the collapse of a large hedge fund in September 1998. After the fallout from the hedge fund collapse had been resolved, the real federal funds rate was restored to an average of about four per cent in 2000. When a new recession appeared in 2001 together with a sustained large collapse in stock market prices, the real federal funds rate was lowered to near zero in the fourth quarter; the rate had averaged zero for 13 consecutive quarters as of March 2005.

Between December 1990 and April 1992 reserve requirements on time and demand deposits were reduced, which helped banks to increase net income. In January 1994 'retail sweep programmes' were introduced. In these programmes, a bank shifts funds from a depositor's transactions account to a synthetic time deposit account in the depositor's name in order avoid reserve requirements, usually without the depositor's knowledge. The Board does not measure the amount of funds swept, except at the time the programme was established. The Board estimated that as of August 1997 required reserves fell by one-third because of these programmes.

In November 1999 President Clinton signed the Financial Services Modernization (Gramm–Leach–Bliley) Act, which reversed the 1933 Glass–Steagall Act's ban on combining commercial and investment banking. The ban had been eroding since 1987, when some large bank holding companies were authorized by the Board to establish subsidiaries that could underwrite state and local government revenue bonds. The new act authorized the establishment of financial holding companies, which were to be regulated by the Board and could engage in an approved list of activities that included commercial banking, insurance, securities underwriting, merchant banking, and complementary financial undertakings. In 2003 there were more than 600 financial holding companies, which resemble the universal banks that exist in other countries.

In December 2002 the Federal Reserve discarded the discount rate as a policy instrument by replacing it with an interest rate on primary credit extended by the discount window that is one per cent above the FOMC target federal funds rate. Primary credits are collateralized loans to banks in sound financial condition.

As the foregoing dramatic institutional changes suggest, the Federal Reserve System is a work in progress. Its set of policy instruments and its dimensions have radically changed. Because of offshore banking facilities and retail sweep accounts, reserve requirement changes are no longer an effective policy instrument. As noted in the preceding paragraph, the discount rate has been discarded as an instrument; it is simply a penalty rate that is related to a bank rate, as is often the practice in other countries. Regulations on the interest rates banks pay on time and savings deposits have been discarded.

Open-market operations are almost the sole policy instrument that can be used to achieve the Board's target nominal and real federal funds interest rates. While the FOMC has been able to control the overnight federal funds rate, the linkage between it and real economic activity is changing. First, the combined holdings of US government securities by foreign central banks have recently exceeded those of Federal Reserve banks. Foreign central bank holdings are partly a result of their efforts to manipulate exchange rates; their holdings are likely to change when FOMC policies change. Second, repurchase agreements and offshore transactions vary considerably over time and their volumes appear to be sensitive to US economic activity. Third, the outstanding stock of securitized mortgage and other debt has been growing rapidly; such debt is a close substitute for US government debt and its amount has real economic effects. Fourth, because of decreasing required reserves and growing offshore holdings of US currency, 89 per cent of Federal Reserve liabilities were in the form of Federal Reserve notes in December 2003; the corresponding share was 34 per cent in 1941, 57 per cent in 1970, and 79 per cent in 1989. In part, the Federal Reserve recently has become an institution for collecting seigniorage from the rest of the world. Finally, over the decade ending in 2003, the share of all credit market assets held by depository institutions in the Federal Reserve's flow of funds accounts fell. In the context of the most recent 13 quarters of a zero real federal funds interest rate, more changes could be expected.

See Also

- ▶ Great Depression
- ▶ Monetary and Fiscal Policy Overview

Bibliography

Bagehot, W. 1873. *Lombard street: A description of the money market*, 1962. Homewood: Richard D. Irwin.

Chandler, L. 1958. *Benjamin strong: Central banker*. Washington, DC: The Brookings Institution.

Chandler, L. 1970. *America's greatest depression 1929–1941*. New York: Harper and Row.

Crabbe, L. 1989. The international gold standard and US monetary policy from World War I to the New Deal. *Federal Reserve Bulletin* 75: 423–440.

Dykes, S. 1989. The establishment and evolution of the Federal Reserve Board: 1913–23. *Federal Reserve Bulletin* 75: 227–243.

Friedman, M., and A. Schwartz. 1963. *A monetary history of the United States, 1867–1960*. Princeton: Princeton University Press.

Goldenweiser, E. 1951. *American monetary policy: A research study for the committee for economic development*. New York: McGraw-Hill.

Greider, W. 1987. *Secrets of the temple: How the Federal Reserve Runs the Country*. New York: Simon and Schuster.

Meltzer, A. 2003. *A history of the Federal Reserve, Volume I (1913–1951)*. Chicago: University of Chicago Press.

Mundell, R. 1961. Flexible exchange rates and employment policy. *Canadian Journal of Economics* 27: 509–517.

Scott, I. Jr. 1957. The availability of credit doctrine: theoretical underpinnings. *Review of Economic Studies* 25(4): 41–48.

Warburg, P. 1930. *The federal reserve system: Its origin and growth*. Vol. 1. New York: Macmillan.

Fel'dman, Grigorii Alexandrovich (1884–1958)

Michael Ellman

Abstract

Fel'dman was one of the founders of the theory of economic growth, the economics of planning and development economics. His contributions were made in the USSR in the late 1920s. He developed a two-sector growth model and showed how different growth rates implied different economic structures. He derived two theorems. He is regarded as the father of the 'heavy industry first' strategy of economic development. A brilliant pioneer, Fel'dman's work was cut short by the Stalinists. Later analysis and international experience revealed a number of limitations of a narrowly Fel'dmanite approach to economic policy.

Keywords
Development economics; Fel'dman, G.; Fel'dman's two theorems; Gosplan; Growth, models of; Kalecki, M.; Planning; Soviet growth record

JEL Classifications
B31

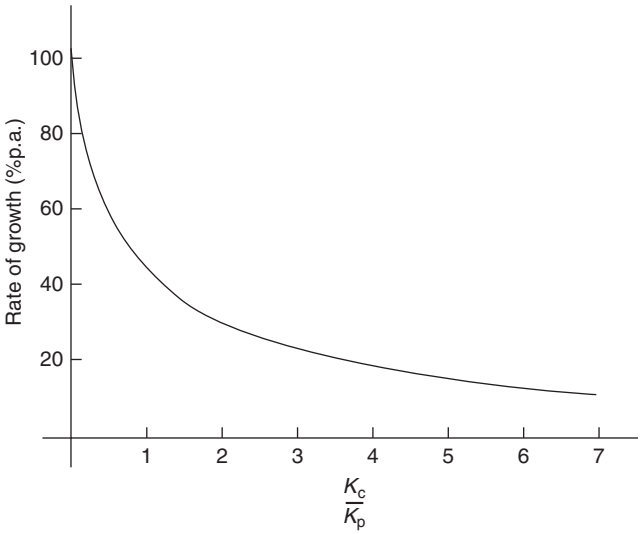
Fel'dman was one of the founders of the theory of economic growth under socialism, the economics of planning and development economics. An electrical engineer by profession, he worked in Gosplan from February 1923 to January 1931. It was in this period that his contribution to economics was made. At first he was in the department analysing and forecasting developments in the world economy (he concentrated on Germany and the USA). His first work on the theory of growth was a comparative study of the structure and dynamics of the US economy in 1850–1925 with projections of the Soviet economy between 1926/1927 and 1940/1941. His most important work ('On the theory of the rates of growth of the national income') was a report to Gosplan's committee for compiling a long-term plan for the development of the national economy of the USSR. It was published in two parts in Gosplan's journal in 1928. A year later Fel'dman published a paper (1929c) which provides a more popular presentation of how to utilize his ideas to calculate long-term plans. The ideas of Fel'dman formed the methodological basis for the preliminary draft of a long-term plan worked out by the committee, then headed by N.A. Kovalevskii. This draft was discussed at meetings of Gosplan's economic research institute in February and March 1930. Apart from this serious discussion, during 1930 Fel'dman came under public attack for his ideas. His reliance on mathematics and his lack of fanaticism did not fit in well with the political fervour of 1930. The concrete numerical work of Fel'dman and Kovalevskii in 1928–1930 was much too optimistic. It treated as feasible entirely unrealizable goals. The attempt to realize them had disastrous effects on the economy. Unfortunately, the political situation in the USSR prevented Fel'dman from publishing anything on economics after 1930. Even when, in 1933, he reverted from the sensitive subject of socialist industrialization to the problems of capitalist growth, his book was not published.

As far as growth theory is concerned, Fel'dman's work was much in advance of contemporary Western work. He developed a two-sector growth model and showed how different growth rates implied different economic structures. He derived two important results, one about the ratios of the capital stocks in the two sectors, the other about the allocation of investment between the two sectors. The first result is that a high rate of growth requires that a high proportion of the capital stock be in the producer-goods sector. This is illustrated in Fig. 1. Fel'dman's second theorem is that, along a steady growth path, investment should be allocated between the sectors in the same proportion as the capital stock. For example, suppose that a 20 per cent rate of growth requires a K_c/K_p of 3.7. Then, to maintain growth at 20 per cent p.a. requires that 3.7/4.7 of annual investment goes to the consumer-goods industries and 1.0/4.7 of annual investment goes to the producer goods industries.

The interrelationship between the two theorems is shown in Table 1, in which Fel'dman explained how any desired growth rate, given the capital–output ratio, determined both the necessary sectoral composition of the capital stock and the sectoral allocation of investment.

Given the capital–output ratio, the higher the K_p/K_c ratio, that is, the greater the proportion of the capital stock in the producer goods sector, and correspondingly the higher the $\Delta K_p/\Delta K_c + \Delta K_p)$ ratio, that is, the greater the proportion of new investment in the producer-goods sector, the higher the rate of growth. With a capital–output ratio of 2.1, to raise the growth rate from 16.2 to 24.3 per cent requires raising the proportion of the capital stock in the producer-goods sector from a third to a half, and the share of investment in the producer-goods sector from a third to a half.

The conclusion Fel'dman drew from his model was that the main tasks of the planners were to regulate the capital–output ratios in the two sectors and the ratio of the capital stock in the producer-goods sector to that in the consumer-goods sector.



Fel'dman, Grigorii Alexandrovich (1884–1958), Fig. 1 Fel'dman's first theorem. Notes: K_c is the capital stock in the consumer goods industry, K_p is the capital stock in the producer goods industry

Fel'dman, Grigorii Alexandrovich (1884–1958), Table 1 Fel'dman's two theorems

$\frac{K_p}{K_c}$	$\frac{dy}{dt}$(in % p.a.) (when $K/Y = 2.1$)	$\frac{\Delta K_p}{\Delta K_c + \Delta K_p}$
0.106	4.6	0.096
0.2	8.1	0.167
0.5	16.2	0.333
1.0	24.3	0.500

For the former task, Fel'dman recommended rationalization and multi-shift working; for the latter, investment in the producer-goods sector.

As far as the economics of planning is concerned, the main lesson to be learned from the Fel'dman model is that the capacity of the capital-goods industry is one of the constraints limiting the rate of growth of an economy. There may well be other constraints, such as foreign exchange, urban real wages or the marketed output of agriculture. (Indeed, it is possible that one or more of these are binding constraints and that the limited capacity of the producer-goods sector is a non-binding constraint.) Economic planning is largely concerned with the removal of constraints to rapid economic growth. Accordingly, a planned process of rapid growth may require that the planners stimulate the rapid development of the producer-goods sector.

As far as development economics is concerned, Fel'dman is important because of the argument in his 1928 paper that 'an increase in the rate of growth of income demands industrialization, heavy industry, machine building, electrification...'. When first formulated, this conclusion struck many economists as counter-intuitive and paradoxical.

Fel'dman's work, as is natural for a pioneer, suffers from serious limitations. As far as the theory of economic growth under socialism is concerned, he was an important early contributor, but his work has to be complemented by Kalecki's (1969) emphasis on the limits of growth and Kornai's (1992, ch. 9) emphasis on the behavioural regularities actually generating the growth process. As for the economics of planning, his arguments have to be complemented by a proper understanding of the role of agriculture, foreign trade and personal consumption and of the danger of an over-accumulation crisis. In development economics, experience in the USSR in the 1930s, India in the 1950s and China in the Maoist period has shown the limitations of a narrowly Fel'dmanite approach.

A brilliant pioneer, Fel'dman's work was ended after only a few years by the Stalinists. In January 1931 Fel'dman was forced out of Gosplan. He seems to have been arrested in 1937 and only released – probably from the Gulag – in 1943, but even then was forbidden to return to Moscow. He was only allowed to return to Moscow in 1953, by which time he was seriously ill.

See Also

- ▶ Development economics
- ▶ Kalecki, Michal (1899–1970)
- ▶ Soviet growth record

Selected Works

1927. Soobrazheniya o strukture i dinamike narodnogo khozyaistvo SSha s 1850 po 1925 g i SSSR s 1926/1927 po 1940/1941 gg [Reflections on the structure and dynamics of the national economy of the USA from 1850 to 1925 and of the USSR from 1926/1927 to 1940/1941]. *Planovoe khozyaistvo* No. 7. Also published as a booklet.

1928. K teorii tempov narodnogo khozyaistva [On the theory of the rates of growth of the national income]. *Planovoe khozyaistvo* Nos. 11 and 12. English translation in *Foundations of Soviet Strategy for Economic Growth*, ed. N. Spulber. Bloomington: Indiana University Press, 1964.

1929a. SSSR i mirovoe khozyaistvo na rubezhe vtorogo goda pyatletki [The USSR and the world economy on the eve of the second year of the five year plan]. *Na planovom fronte* No. 2.

1929b. O limitakh industrializatsii [On the limits of industrialization]. *Planovoe khozyaistvo* No. 2.

1929c. Analiticheskii metod postroeniya perspektivnykh planov [An analytical method for constructing perspective plans]. *Planovoe khozyaistvo* No. 12.

1930. Problemy electrifikatsii na novom etape [Problems of electrification at a new stage]. In *Na novom etape sotsialisticheskogo stroitel'stva,* vol. 1. Moscow: Gos. planovoe khozyaistvennoe izd-vo.

Bibliography

Soviet Evaluations

Planovoe khozyaistvo. 1930. Report of the discussion in Gosplan's economic research institute of February and March 1930. No. 3, 117–211.

Vainshtein, A. and G. Khanin 1968. Pamyati vydayushchegocya sovetskogo ekonomista-matematika G.A. Fel'dmana [In memory of the outstanding Soviet mathematical economist G.A. Fel'dman]. *Ekonomika i matematicheskie melody* 4(2).

English-Language Works

Allen, R. 2003. The development problem in the 1920s. Ch. 3. In *Farm to factory*. Princeton: Princeton University Press.

Chng, M. 1980. Dobb and the Marx-Fel'dman model. *Cambridge Journal of Economics* 4: 393–400.

Dobb, M. 1967. *The question of 'Investment priority for heavy industry'*, Ch. 4 of Papers on Capitalism, Development and Planning. London: Routledge.

Domar, E. 1957. A soviet model of growth. In *Essays in the theory of economic growth*. New York: Oxford University Press.

Erlich, A. 1978. Dobb and the Marx-Fel'dman model: a problem in Soviet economic strategy. *Cambridge Journal of Economics* 2: 203–214.

Kalecki, M. 1969. *Introduction to the theory of growth in a socialist economy.* Oxford: Blackwell.

Kornai, J. 1992. *The socialist system: The political economy of communism*. Oxford: Oxford University Press.

Mahalanobis, P. 1953. Some observations on the process of growth of national income. *Sankhya* 12: 307–312.

Mahalanobis, P. 1955. The approach of operational research to planning in India. *Sankhya* 16: 3–130.

Sen, A., and K. Raj. 1961. Alternative patterns of growth under conditions of stagnant export earnings. *Oxford Economic Papers* 13: 43–52.

Tinbergen, J., and H. Bos. 1962. *Mathematical models of economic growth*. New York: McGraw.

Fellner, William John (1905–1983)

Irma Adelman

Fellner was born in Budapest and received his PhD at the University of Berlin. In 1938 he moved to the United States and taught at Berkeley (1939–52) and Yale (1952–73). He was President of the American Economic Association (1969) and a member of the Council of Economic Advisers (1973–5).

His major contributions were to macroeconomic theory and policy. Those who, like myself, were fortunate enough to know him came to worship him because of his combination of nobility of spirit, profundity, subtlety, humility, deep culture

and inherent humanity. His writings were shaped by all these qualities as well as by his formative experiences in interwar Europe. He was a liberal of the old school – a humanist and an anti-authoritarian. He had been traumatized by the German hyperinflation, the mass unemployment of the Great Depression, and by the Nazi totalitarianism which ensued. His teachings were committed to avoiding a repetition.

Upon re-reading his *Monetary Policies and Full Employment* (1946), one is struck by how far ahead of his time he was. A limited Keynesian, he advocated policies aimed at avoiding severe recessions and allowing small ones to run their course because he foresaw that an unconditional guarantee of full employment would lead monopolistic groups of industrialists and workers to constantly raise wages and prices and reduce quality. This is precisely what happened in the industrial countries between 1965 and 1973. The result would be that an unconditional full employment guarantee would require government controls on wages and prices and would ultimately result in a severe abrogation of both liberty and market efficiency. He argued that growth and cycles are interdependent as are price stability and employment. He emphasized the role of uncertainty, expectations and credibility of government-policy commitments. He foreshadowed both a subtler version of supply-side economics and of rational expectations. With respect to supply-side economics, he argued that fiscal expansionism should be limited to counteracting severe recessions only and that otherwise a combination of credit policies, price- cost policies, and tax policies aimed at increasing the level of private activity would be preferable. He argued that price-wage expectations are critical since uncertainty could defeat Keynesian policies.

Thirty years later (Fellner 1976), he amplified this theme. He suggested that dynamic macroeconomic equilibrium requires not only that savings-investment decisions be validated but also that price expectations be close to actual price levels. He qualified the now usual rational expectations model with the view that public predictions of government reactions are probabilistic; credibility is, therefore, critical. But government should not passively validate just any expectations. Rather, a major policy aim should be to create an environment of restraint which leads to stable rather than explosive expectations. Thus, he cast government in the same role as himself – that of a wise teacher.

Selected Works

1946. *Monetary policies and full employment*. Berkeley: University of California Press.
1949. *Competition among the few*. New York: Alfred A. Knopf.
1955. *Trends and cycles in economic activity*. New York: Holt, Rinehart & Winston.
1960. *Emergence and content of modern economic analysis*. New York: McGraw-Hill.
1965. *Probability and profit*. Homewood.: Richard D. Irwin.
1976. *Towards a reconstruction of macroeconomics*. Washington, DC: American Enterprise Institute.

Female Labour Force Participation: Persistence and Evolution

Paola Giuliano

Abstract

This article explores the relevance of deep historical forces that have influenced the historical gender division of labour and the perception of women's roles in society more generally. In particular, we will review how different types of subsistence activity in the ancient past – such as hunting and gathering and various types of agricultural technology – and geography and language can affect the role of women and their relative bargaining positions up to modern times. Finally, we will review the relevance of mechanisms such as learning, in contrast to deep historical forces, to explain the evolution of female labour force participation.

Keywords
Beliefs; Culture; Female labour force participation; Gender roles; Values

JEL Classification
D03; J16; N30

Social attitudes toward women and their role in society show remarkable differences across countries, including those with similar institutions or economic development; in some countries, they have also changed dramatically in a relatively short time.

The economics literature initially explained differences in female labour force participation by looking at standard economic variables such as the level of development, women's education, fertility and marriage/divorce prospects and the expansion of the service sector (see Goldin (1990) for a review). Some scholars have emphasised the role played by market prices, such as the decline in childcare costs (Attanasio et al. 2008), and by technological factors such as the invention of baby formula (Albanesi and Olivetti 2014).

A more recent literature has argued that differences in female labour force participation across countries could reflect underlying cultural values and beliefs, which tend to be transmitted from parents to children and to stay fairly stable over time. This article will review the literature on the relevance of culture in the determination of female labour force participation and especially on the long-term historical origins of these differences, which will help us understand their persistence. We will also look at research emphasising a change in the bargaining power of women inside the married couple which helps explain the dramatic increase in female labour force participation in many countries over the last century. Concluding remarks will discuss directions for further research.

Persistence in Female Labour Force Participation

In 2000, the share of women aged 15–64 in the labour force ranged from 16% in Pakistan to 90.5% in Burundi. Traditional economic interpretations having proven insufficient to explain these differences, a recent strand of literature has emphasised the role of culture. In an important contribution, Fernandez and Fogli (2009) show that female labour force participation amongst second-generation immigrants in the USA is very strongly correlated with female labour force participation in the country of origin. This evidence is relevant to explain the importance of culture, because migrant women of various origins are all observed in the same institutional and labour market environment. (The authors chose second-generation immigrants because the problem of selection and disruption due to migration is less relevant for them than for first-generation immigrants.)

Although this evidence clearly shows that culture matters, little is known of the historical origin of these cultural differences. In this section, we will look at three important long-term historical determinants of gender roles: agricultural technology, language and geography.

Differences in Historical Agricultural Technologies

Alesina et al. (2013) study the historical persistence of differences in female labour force participation. The hypothesis for their empirical analysis comes from the seminal work of Ester Boserup (1970), in which she argued that differences in the role of women in societies originate in different types of agricultural technologies, particularly the differences between shifting and plough agriculture. Shifting agriculture, which uses hand-held tools such as the hoe and the digging stick, is labour-intensive, with women actively participating in farm work, while plough agriculture is more capital-intensive, using the plough to prepare the soil. Unlike the hoe or digging stick, the plough requires significant upper-body strength, grip strength and bursts of power to either pull the plough or control the animal that pulls it. Farming with the plough is also less compatible with childcare, which is almost always the responsibility of women. As a result, men tended to specialise in agricultural work outside the home, while women specialised

in activities within the home. In turn, this division of labour generated different norms about the appropriate role of women in society. Societies characterised by plough agriculture developed the belief that the natural place for women is in the home. This belief tends to persist even if the economy moves out of agriculture, affecting the participation of women in activities performed outside the home, including market employment, entrepreneurship and politics.

The authors start their analysis by documenting a very strong negative correlation between traditional use of the plough and female participation in agriculture in pre-industrial societies, using the Ethnographic Atlas, a dataset assembled by George Peter Murdock in 1967 and containing ethnographic information for 1,265 ethnic groups covering the whole world. To investigate whether plough-based agriculture correlates with lower female participation in all agricultural tasks or only in a few (such as soil preparation), the authors report results on specific activities carried out in the field or outside the home: land clearance, soil preparation, planting, crop tending, harvesting, caring for small and large animals, milking, cooking, fuel gathering, water fetching, burden carrying, handicraft production and trading. Their empirical analysis carefully controls for all the other variables that could be correlated with plough use and gender roles: the presence of large domesticated animals, a measure of economic development, the fraction of land where the ethnic group lives defined as tropical or subtropical, and the fraction of land that is defined as overall suitable for agriculture. Overall, the authors find that plough use is associated with less female participation in all agricultural tasks, with the largest declines in soil preparation, planting, crop tending and burden carrying. But they find that plough use tends not to be significantly correlated with female participation in other activities. This interpretation of the correlations is fully consistent with Boserup's hypothesis.

After looking at the correlation between agricultural technology and female participation in agriculture in pre-industrial societies, Alesina et al. (2013) study whether differences in agriculture technologies still have an impact on female labour force participation today. The existence of a correlation between female labour force participation in agriculture and agricultural technology in the past does not necessarily imply that differences in historical agriculture technologies affect female labour force participation today. Goldin and Sokoloff (1984), for example, document that within the northeastern USA the low relative productivity of women and children in agriculture (and their low participation in this sector) allowed them to participate actively in the manufacturing sector. In this setting, initial female labour force participation in agriculture is inversely related to subsequent participation in manufacturing, showing a lack of continuity of female labour force participation over time as industrialisation occurred. An interpretation based on social norms could, however, help explain the long-term persistence.

To show long-term persistence, Alesina et al. (2013) look at differences in female labour force participation, but also at beliefs about the role of women in society in 2000.

To analyse contemporary female labour force participation, they match ethnographic data to current populations using the global distribution of 7,612 language groups from the 15th edition of the *Ethnologue* and the global distribution of population densities from the 2000 *Landscan* database, generating a measure of the fraction of a country's ancestors who traditionally engaged in plough agriculture.

At the country level, the authors look at differences in female labour force participation and also at two other measures that could reflect cultural attitudes and beliefs about the role of women in society: a measure of entrepreneurship (given by the share of firms with a woman among the principal owners) and the presence of women in national politics (given by the proportion of parliamentary seats held by women). In countries with a tradition of plough use, women are less likely to participate in the labour market, own firms and participate in national politics.

Along the lines of Alesina et al. (2013), Hansen et al. (2012) hypothesise that societies with long histories of agriculture have less equality in

gender roles as a consequence of more patriarchal values and beliefs regarding the proper role of women in society. Their research is motivated by the idea that patriarchy originated in the Neolithic Revolution – the prehistoric transition from a hunter-gatherer society to an agricultural one – and that patriarchal values and beliefs have persisted and become more ingrained in countries with long histories of agriculture. Agricultural societies were more gender-biased than hunter-gatherer societies. Population growth and land scarcity made cultivation of food more labour-intensive, which created 'a premium on male brawn in plowing and other heavy farm work' (Iversen and Rusenbluth 2010). This led to a division of labour within the family, where the man used his physical strength in food production and the woman took care of child rearing, cooking and other family-related duties. This increased the male's bargaining power within the family, which, over generations, translated into norms and behaviour which shaped cultural beliefs on gender roles.

Using a world sample, a European regional sample and a sample of children of immigrants living in the USA, the authors find a negative association between the number of years that a country had been an agrarian society in 1500 CE and measures of gender equality, including female labour force participation, number of years since women gained suffrage and percentage of seats in parliament held by women.

Language

Another interesting aspect of long-term persistence in gender roles is the relation between grammatical gender-marking and female participation in the labour market, the credit market, land ownership and politics (Gay et al. 2013). The grammatical features of a language are inherited from the distant past and the gender system is one of the most stable linguistic features, surviving for thousands of years. Gay et al. (2013) broadly follow Whorf (1956): 'We are inclined to think of language simply as a technique of expression, and not to realize that language first of all is a classification and arrangement of the stream of sensory experience which results in a certain world-order, a certain segment of the world that is easily expressible by the type of symbolic means that language employs'.

In linguistics, a grammatical gender system is defined as a set of rules for agreement that depends on nouns of different types. These are normally based on biological sex, but can also be based on social constructs, such as age or social status. Gay et al. (2013) rely on the *World Atlas of Linguistic Structures*, the most comprehensive data source of grammatical structures, and use four very stable grammatical variables related to gender: the number of genders in the language, whether the gender system is sex-based, rules for gender assignment and gender distinctions in pronouns. The authors construct the Gender Intensity Index by summing these features for the most commonly spoken language in a country.

Using cross-country and individual-level data, they find that women speaking languages that more pervasively mark gender distinctions are less likely to participate in economic and political activities and more likely to encounter barriers in their access to land and credit. The authors also investigate a sample of migrants living in the USA – that is, all facing the same institutional and labour-market environment – and find consistent results.

Geography

A long-term determinant of differences in gender roles can be found in geography. In a fascinating paper, Carranza (2012), having pointed out that soil texture, which varies exogenously, determines the workability of the soil and the technology used in land preparation, uses this as a lens to look at differences in female labour force participation in India. Deep tillage of land reduces the need for transplanting, fertilising and weeding, which are typically performed by women (Basant 1987). In areas where deep tillage is required, the lower demand for female labour relative to the demand for male labour is expected to have a negative impact on the perceived relative value of girls to a household (Boserup 1970).

Carranza (2012) finds that soil texture explains a large part of the variation in women's relative participation in agriculture. The author also goes

further and examines the impact of geography on infant sex ratio, perhaps the most extreme indicator of gender-based discrimination. Because relatively smaller female labour contributions in loamy areas make girls relatively more costly, the ratio of girls to boys will be negatively related to the difference between the fractions of loamy and clayey soils. Sex ratios and female labour force participation in India show a large geographical heterogeneity, even within the same state and cultural region (Dyson and Moore 1983; Agnihotri 1996). These differences within the same state are not driven by alternative mechanisms, including cultural, social, economic or policy variables.

Carranza (2012) estimates that soil texture explains 62% of the within-state variation in female agricultural labor force participation and 70% of the variation in the sex ratio for zero- to six-year-olds. A 10 percentage point greater fraction of loamy soils relative to clayey soils is associated with a 5.1% lower share of female agricultural labourers and a 2.7% lower ratio of female to male children. The relationship between soil texture, relative female labor-force participation, and the ratio of female to male children did not change significantly between 1961 and 2001.

Alesina et al. (2013) also examine the effect of geography on female labour force participation. They run both instrumental variables regressions and reducedform regressions using the suitability of the soil for crops that do or do not benefit from the use of the plough. The primary benefit of the plough is that you can cultivate a given amount of land more quickly and thus you can cultivate more land in a given amount of time. This capability is more advantageous for crops that require specific planting conditions that occur during narrow windows of time or for crops that require more land to cultivate a given amount of calories. The benefit of the plough is reduced or eliminated for crops grown in swampy, sloped, rocky or shallow soils, where it is less efficient or impossible to use. Taking these factors into consideration, crops can be classified into 'plough-positive crops' – those such as wheat, teff, barley and rye whose cultivation benefits greatly from the use of the plough – and 'plough-negative crops' – those such as sorghum, maize, millet, roots, tubers and tree crops, whose cultivation benefits less from the use of the plough (Pryor 1985).

The authors' estimates show that the adoption of the plough is positively correlated with an environment suitable for plough-positive crops, but not with an environment suitable for plough-negative crops. In a different specification, the authors look directly at the relationship between crop suitabilities and current gender roles. They find that having an ancestral environment that was more suitable to plough-positive crops is always associated with less equal gender roles today, while an environment more suitable to plough-negative crops is generally associated with more equal gender roles today.

Historical Changes in Female Labour Force Participation

A unifying interpretation for the historical change in labour force participation among married women supposes that a working wife has become more attractive to married couples (differences in female labor force participations are more pronounced among married women). In the previous section, we analysed the long-term determinants of gender roles. In this section we will review the factors that determined their evolution. The explanation in the literature has centred on four factors: women having more marital bargaining power because they spend less time on household chores (Greenwood et al. 2005), a changing social atmosphere (Fernandez et al. 2004), the introduction of the contraceptive pill (Goldin and Katz 2002) and learning about the effects of female labour force participation (Fogli and Veldkamp 2011; Fernandez forthcoming).

The Adoption of Household Technology

According to Greenwood and his co-authors (2005), married women could not enter the labour force until housework had become less time-consuming. Specifically, the authors focused on the widespread adoption of household technologies – such as washing machines, vacuum cleaners and dishwashers – that greatly

reduced the time needed to do housework. The authors consider a household in which the husband always works in the labour market and the wife always does the housework. A decline in the price of the technology (which is at the origin of its widespread diffusion) had a large impact on women's labour force participation: more than half of the increase in women's labour force participation was due to labour-saving technology. For comparison, only one-fifth of the increase was directly due to the decline in the gender wage gap. The main conclusion is that a better outside option could encourage women to join the labour force only after the technology to free their time had appeared.

Changing Social Norms

Fernandez et al. (2004) hypothesised that men with working mothers were more likely to have working wives. A son's preference to marry a woman who works may have been influenced by having a working mother. Also, a working mother would be motivated to make her son more productive with household chores, which would later allow his wife more time for work outside the home.

The authors find that the probability that a married woman worked full-time was 32 percentage points higher if her husband's mother had worked for at least one year when he was young. To rule out the possibility that the husband's behaviour is determined by assortative mating of individuals whose mothers had worked, the authors look at whether a mother's decision to work also affects her daughter's decision to work. Surprisingly, the wife's decision to work was unaffected by her own mother's labour force status.

The Pill

Goldin and Katz (2002) focus their attention on the birth control pill. According to the authors, the pill caused an increase in female labour force participation because it changed the age at which women married and became pregnant. Goldin and Katz (2002) argue that the pill's availability to unmarried college-aged women increased their career investment and, hence, their long-term labour force participation. Without the pill, young women who wanted professional careers would have to either practice abstinence or run the risk of pregnancy. The pill, in contrast, meant that women did not have to choose one or the other, which lowered the cost of delaying marriage and investing in a long-term career.

Marital decisions across groups reflect the effect of the pill's availability on the workforce decisions of single young women. The proportion of female college graduates born in 1950–54 who were married by age 23 declined by 8.7 percentage points, compared with those born in 1940–49. Access to the pill by age 17 lowered the fraction of married women by 3.2 percentage points (37% of the total decline).

Goldin and Katz (2002) also look at long-term career investments, estimating an increase, between 1970 and 1990, of five percentage points in the share of 30- to 49-year-old women in professional occupations. Approximately 1.7 percentage points (one-third of that increase) can be attributed to increased pill use. The effect is even bigger for the share of college women who became doctors and lawyers: of the total increase of 1.7 percentage points, increased use of the pill explains 1.2 percentage points (three-fourth of the total).

Learning

Two recent papers emphasise the role of learning in the transition from a low to a high level of female labour force participation in the USA.

Fogli and Veldkamp (2011) develop a model in which women learn about the effects of maternal employment on children by observing nearby employed women. When few women participate in the labour force, information is scarce and participation rises slowly. As information accumulates in some regions, the effects of maternal employment become less uncertain and more women in that region participate. Learning accelerates, labour force participation rises more quickly and regional participation diverges. Eventually, information diffuses throughout the economy, beliefs converge to the truth, participation flattens out and regions become more similar again. This model generates changes in female labour force participation that are geographically heterogeneous, locally correlated and smooth in

the aggregate, corresponding to the trends in historical female labour force participation data.

Fernandez (forthcoming) develops a model in which labour force participation by married women and cultural beliefs about the role of women in society evolve jointly. The basic idea is that the probability that individuals assign to different views of the long-term consequences of married women working is updated in a Bayesian fashion as new information endogenously becomes available. Married women compare the benefits of increased consumption from labour earnings with the expected utility cost of working. This cost was at first unknown and women's beliefs about it evolved endogenously over time in a Bayesian fashion. A model with these features, calibrated to key statistics from the twentieth century, generates a time-trend of labour force participation by married women that corresponds to its historical evolution in the USA over the last 120 years.

Concluding Remarks

Differences in female labour force participation have long been remarkably stable. At the same time, female labour force participation has increased quickly in several countries. We review the historical origins of the observed persistence and also study the most recent factors that made a working wife more attractive to married couples and therefore implied an increase in female labour force participation. Several questions remain open: what determines the speed of the evolution of gender roles? When do gender role differences persist or not? What factors affect their persistence?

The literature has so far focused on documenting historical and cultural persistence, yet this persistence remains poorly understood. Various reasons could explain it: underlying cultural traits may be reinforced by policies, laws and institutions which affect the benefits of beliefs about gender inequality. A society with traditional beliefs about gender inequality may, for example, perpetuate these beliefs by institutionalising unequal property and voting rights. Beliefs about gender inequality may also cause a society to specialise in capital-intensive industries, which in turn decreases the relative cost of those gender-inequality norms, which in turn helps perpetuate them. More research should be done to understand these interactions.

Most of the papers in the literature also try to examine an event in isolation from other events, except possibly to account for other covariates. However, the evolution of gender roles is much more complex and highly nonlinear. Understanding the evolution of female labour force participation will depend on obtaining a chronology of both cultural and institutional changes and examining the interrelationships between them.

F

See Also

- ▶ Cultural Transmission
- ▶ Culture and Economics
- ▶ Gender Roles and Division of Labour
- ▶ Women's Work and Wages

Bibliography

Agnihotri, S. 1996. Juvenile sex ratios in India: A disaggregated analysis. *Economic and Political Weekly* 31(52): 3369–3382.

Albanesi, S., and C. Olivetti. 2014. *Gender roles and medical progress*. Mimeo: Boston University.

Alesina, A., P. Giuliano, and N. Nunn. 2013. On the origins of gender roles: Women and the plough. *Quarterly Journal of Economics* 128(2): 469–530.

Attanasio, O., H. Low, and V. Sanchez-Marcos. 2008. Explaining changes in female labor supply in a life-cycle model. *American Economic Review* 98: 1517–1552.

Basant, R. 1987. Agricultural technology and employment in India: A survey of recent research. *Economic and Political Weekly* 22(32): 1348–1364.

Boserup, E. 1970. *Woman's role in economic development*. London: George Allen and Unwin.

Carranza, E. 2012. *Soil endowments, production technologies and missing women in India*. The World Bank Policy Research Working Paper 5974.

Dyson, T., and M. Moore. 1983. On kinship structure, female autonomy and demographic behavior in India. *Population and Development Review* 9(1): 35–60.

Fernandez, R. Forthcoming. Cultural change as learning: The evolution of female labor force participation over a century. *American Economic Review*.

Fernandez, R., and A. Fogli. 2009. Culture: An empirical investigation of beliefs, work and fertility. *American Economic Journal: Macroeconomics* 1(1): 146–167.

Fernandez, R., A. Fogli, and C. Olivetti. 2004. Mothers and sons: Preference formation and female labor force dynamics. *Quarterly Journal of Economics* 119(4): 1249–1299.
Fogli, A., and L. Veldkamp. 2011. Nature or nurture? Learning and the geographic of female labor force participation. *Econometrica* 79(4): 1103–1138.
Gay, V., E. Santacreu-Vasut, and A. Shoham. 2013. *The grammatical origins of gender roles*. Berkeley Economic History Laboratory, WP2013-03.
Goldin, C. 1990. *Understanding the gender gap: An economic history of American women*. New York: Oxford University Press.
Goldin, C., and L. Katz. 2002. The power of the pill: Oral contraceptives and women's career and marriage decisions. *Journal of Political Economy* 110(4): 730–770.
Goldin, C., and K. Sokoloff. 1984. The relative productivity hypothesis of industrialization: The American case, 1820 to 1850. *Quarterly Journal of Economics* 99(3): 461–487.
Greenwood, J., A. Seshadri, and M. Yorukoglu. 2005. Engines of liberation. *Review of Economic Studies* 72(1): 109–133.
Hansen, C.W., P. Jensen, and C. Skovsgaard. 2012. *Modern gender roles and agricultural history: The Neolithic inheritance*. Mimeo: University of Southern Denmark.
Iversen, T., and F. Rosenbluth. 2010. *Women, work and politics: The political economy of gender inequality.* New Haven: Yale University Press.
Murdock, G. 1967. *Ethnographic atlas*. Pittsburgh: University of Pittsburgh.
Pryor, F. 1985. The invention of the plow. *Comparative Studies in Society and History* 27(4): 727–743.
Whorf, B.L. 1956. The punctual and segmentative aspects of verbs in Hopi. In *Language, thought and reality: Selected writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press.

Feminist Economics

Julie A. Nelson

Abstract

Feminist economics is a field that includes both studies of gender roles in the economy from a liberatory perspective and critical work directed at biases in the economics discipline. It challenges economic analyses that treat women as invisible, or that serve to reinforce situations oppressive to women, and develops innovative research designed to overcome these failings. Feminist economics points out how subjective biases concerning acceptable topics and methods have compromised the reliability of economics research. Topics addressed include the economics of households, labour markets, care, development, the macroeconomy, national budgets, and the history, philosophy, methodology, and teaching of economics.

Keywords

Sen, A; American Economic Association; Bargaining models; Caring work; Children; Data quality; Development; Divorce; Ecology; Economic man; Equal rights movement; Feminist economics; Fertility; Gender; Gender equity; Gender roles; Health care; Household production; Household labour; Institutional economics; International Association for Feminist Economics; Innate differences; Intrahousehold distribution; Labour market discrimination; Marriage; Masculinist bias; Methodology of economics; Occupational segregation; Opportunity cost method; Post Keynesian economics; Postmodernism; Race; Radical economics; Replacement cost method; Social economics; Structural adjustment; Subcontracting; Trade liberalization; Unpaid work; Women's economic status

JEL Classifications

B31

Feminist economics is a field that includes both studies of gender roles in the economy from a liberatory perspective and critical work directed at biases in the content and methodology of the economics discipline. It challenges economic analyses that treat women as invisible or that serve to reinforce situations oppressive to women, and develops innovative research designed to overcome these failings. Feminist economics points out how subjective biases concerning acceptable topics and methods have compromised the reliability and objectivity of economics research, and explores more adequate alternatives.

The Origins of Feminist Economics

Feminist economics in its contemporary form began in the 1970s in response to the prevailing pattern of labour market and household studies. Up until the 1960s, women and women's traditional activities had been subsumed into the 'black box' of the household within neoclassical economics. Neoclassical theory had been defined as the study of choices made in markets by rational, autonomous actors. A household was generally understood to be represented by its male 'head', whose preferences, it was assumed, determined household labour supply and consumption decisions. The household was assumed to enjoy a single utility level, and activities within the household were classified as 'leisure'. Studies of paid labour generally focused on men only, and household production was (and is still) excluded from national accounts. Women, women's traditional activities, and the well-being of women and children were invisible.

During the 1960s, issues of labour market discrimination by race and sex began to be debated. The idea that household activities might include unpaid work as well as leisure also gained ground. The New Home Economics school sought to extend rational choice theory to intra-household decisions. Often, however, work by economists on these issues simply defended traditional sex roles in the family, women's segregation into a narrow range of paid occupations, and women's lesser earnings in the paid labour market. In general, neoclassical economists of the time argued that the prevailing patterns resulted from rational choices, with variations between men and women due only to presumably innate differences between men and women in tastes and abilities, often expressed in different choices about human capital formation. As well, circular reasoning was used: women's lesser market earnings were used to explain their specialization in household work, and women's household responsibilities were used to justify their lesser market earnings. While these works recognized women's existence, they were not feminist in that they served to rationalize rather than explore and question women's assignment to second- class status and financial dependency.

A key distinction feminist economists make is between sex, understood as the biological difference between males and females, and *gender*, the social beliefs that society constructs on the basis of sex. While traditional economists saw household and labour market outcomes as reflecting only sex differences, feminist economists raised the question of how much these outcomes might, instead, reflect misleading stereotypes and rigid social constraints. Some works called into question, for example, the ideas that specialization in household work would be an optimizing choice for a woman (given rising divorce rates) or that it would necessarily yield greater household well-being than other, more egalitarian, arrangements (Ferber and Birnbaum 1977; all references given in this article are examples from larger literatures). Others emphasized the role of discrimination in limiting women's labour market opportunities (Bergmann 1974) or the interplay of household and workplace power relations (Hartmann 1976).

In actuality, as the equal rights movements of the 1960s and 1970s loosened many of the legal restrictions and social norms that had artificially narrowed women's educational and job choices in a number of countries, women moved increasingly into the labour market and into formerly all-male occupations. Surveys of women's economic history, economic status, and progress towards gaining economic equality have since been undertaken for many countries and regions, along with surveys of policies related to gender equity. Recognition of the importance of social beliefs and power structures in creating gendered economic outcomes has remained a hallmark of feminist economics.

The Critique of Mainstream Economics

While feminist were originally dissatisfied with mainstream economic scholarship because it neglected and distorted women's experiences, by the late 1980s feminists were also advancing a more thoroughgoing critique. Many feminist economists were finding that traditional formal choice-theoretic modelling and a narrow focus

on mathematical and econometric methods were a Procrustean bed when it came to analysing phenomena characterized by connection to others, tradition, and relations of domination. Feminists began to raise questions about the mainstream definition of economics, its central image of 'economic man' and the exclusive use of a particular set of methodological tools.

Essays on this theme were brought together in a 1993 volume, *Beyond Economic Man: Feminist Theory and Economics* (Ferber and Nelson 1993). In this volume it was suggested that economics be defined by a concern with the provisioning of life in all spheres where this occurs rather than only in markets. Investigations were undertaken into how a particular set of professional values, emphasizing culturally masculine-associated factors such as autonomy, separation, and abstraction, had come to take precedence over culturally feminine-associated factors such as interdependence, connection, and concreteness. The contributors argued that, rather than taking the former as a sign of 'rigour' in the discipline, the truncation of methods created by masculinist bias had weakened the discipline's ability to explain real-world phenomena. Questions were raised about mainstream economics not because it was too objective but because it was not objective enough.

A conference held in Amsterdam in 1993 further developed this theme, and contributed innovative discussions on economic methodology (Kuiper and Sap 1995). While many feminist economists continue to make use of traditional mainstream tools, on the whole the field has come to be characterized by the inclusion of a broader range of concepts and methods. Theories of human behaviour that include a balance between individuality and relationship, autonomy and dependence, and reason and emotion are being developed (Ferber and Nelson 1993, 2003). The use of historical studies, case studies, interviews and other qualitative data, as well as greater attention to issues such as data quality and replication in quantitative work, are being explored (Bergmann 1989; Nelson 1995). Feminist economists tend to find that such serious efforts to create and promote more adequate forms of economic practice lead to new insights across the board, whether or not the topic being studied is explicitly gender-related.

The Formation of a Field

With the publication of a number of books and articles, and gatherings at early conferences, feminist economics coalesced into an organized field in the early 1990s. The International Association for Feminist Economics was formed in 1992, and its journal, *Feminist Economics*, commenced publication a few years later (Strassmann 1995). The field was first described in a journal of the American Economic Association in 1995 (Nelson 1995), an encyclopedia of feminist economics was published in 1999 (Peterson and Lewis 1999), and a review of developments during the first ten years of feminist economics was published in 2003 (Ferber and Nelson 2003).

International and wide-ranging in scope, feminist economics now includes work on a number of subjects, including topics in microeconomics, macroeconomics, history, and philosophy.

Labour, Households and Care

True to its roots, feminist economics continues to develop analyses of gender roles in labour markets and households. Many studies of women's paid labour supply, labour market discrimination, and the origins of occupational segregation have been undertaken. Some feminists make use of mainstream theories or econometric models to examine the wage gap between men and women and its possible explanations. Other feminist economists raise questions about the ability of such tools, used alone, to shed light on the underlying causes of inequality, and encourage increased investigation into the social, political and institutional structures of gender and labour markets (Bergmann 1989; Rubery 1998; Figart et al. 2002).

Studies of unpaid work within households have sought to obtain quantitative measures of this labour and to increase the attention paid to unpaid work in the design of policies

(Waring 1988; Ironmonger 1996). The issue of valuing this work remains controversial among feminists. Some feminists endorse the use of replacement cost or opportunity cost methods of assigning dollar values to unpaid household labour. Others argue that these methods lead to understatement because the wages used in such imputations have been kept artificially low by discrimination. Still others believe that this issue serves to draw attention away from women's lack of access to real money and power.

Issues of intra-household distribution and decision-making have been investigated by many feminist economists. The dramatic effect of skewed intrahousehold distribution by sex in countries such as China, India and Pakistan has been brought to public attention (Sen 1990). Bargaining models (McElroy and Horney 1981) have been developed as one way of bringing women's agency within households to the fore. Issues concerning marriage, divorce, fertility and the wellbeing of children have been investigated from feminist perspectives. A number of feminist economists go beyond choice-theory-based bargaining models to examine legal, social, and psychological issues related to intra-household decision-making and well-being (Sen 1984, ch. 16; Agarwal 1997; Wheelock et al. 2003).

Much of women's traditional work in sex-segregated occupations (such as nursing and childcare) and within households can be described as 'caring work'. Caring work presents a challenge to mainstream economics since the traditional image of 'economic man' is of an autonomous, self-interested individual who neither requires care nor has any inclination to provide it. The conceptual and empirical study of work with dependency, emotional or other-regarding components has recently become a field of active investigation for many feminist economists (Folbre 1994; Himmelweit 1999; Folbre and Nelson 2000; Bettio and Plantenga 2004).

Feminist economists have developed critiques of theories and policies that assume that economic agents are unencumbered prime-age workers, and that delve into the economic problems of elderly women, parents of young children, and lone mothers who are faced with simultaneous responsibilities for income generation and family care (MacDonald 1998; Albelda et al. 2004).

Development, Macroeconomics and National Budgets

Feminist economists have also made innovations in the analysis of national and global economies. Studies of the effects of including unpaid production in GDP (Wagman and Folbre 1996) and the analysis of government budgets according to their effects on gender equity (Budlender et al. 2002) have become well-developed fields.

Feminist economists have challenged the definition of economic development in terms of industrialization and GDP growth, drawing attention instead to issues of growth in human well-being and capabilities (Elson 1991; Benería 2003; Agarwal et al. 2003). Many have studied the changes in women's status that have come about during transitions from socialism and during other forms of macroeconomic restructuring (Aslanbeigui et al. 1994).

The effects of macroeconomic policies of structural adjustment and the liberalization of global trade and finance have been looked at from a feminist point of view (Çagatay et al. 1995; Grown et al. 2000). For example, programmes that prescribe macroeconomic belt-tightening through cutbacks in health care often have their most immediate impact on women, as women are expected take on, unpaid, the work of providing services no longer provided by governments. Men and women may also be affected differently, depending on the degree to which they work in subsistence or traded sectors. Women's employment in subcontracting firms has also received considerable attention (Kabeer 2000; Balakrishnan 2002).

History, Philosophy and Teaching

As well as investigating the history of women's economic activities (Humphries 1990) and the history of economic thought in relation to women (Folbre 1991; Pujol 1992), feminist economists

have looked at the history of women and feminists within the economics discipline itself (Dimand et al. 1995). The most recent national studies indicate that women are still under-represented in the top ranks of academic economics (Booth et al. 2000), receiving tenure less frequently than men even when factors such as publications and family are controlled for (Ginther and Kahn 2004). Sexual harassment, sex discrimination, and inhospitable environments are among the barriers yet to be overcome in some departments and universities (Ginther and Kahn 2004).

Feminist economists have also engaged in philosophical discussion concerning the epistemological and methodological foundations of economics in dialogue with postmodernist, postcolonialist, critical realist and other perspectives (Barker and Kuiper 2003). Feminists have also explored comparisons of aims and methods with various heterodox schools of economics including institutionalist economics (Waller and Jennings 1990), social economics (Emami 1993), radical economics (Matthaei, and Post Keynesian economics (Danby 2004).

Regarding the teaching of economics, feminist economists have investigated how the content of economics courses can be made less biased concerning women (Feiner 2004), how courses can be enriched by feminist re-evaluation of theories and methods, and how pedagogy can be adapted to better reach students with diverse backgrounds and learning styles (Shackelford 1992; Aerni and McGoldrick 1999).

Feminism and Other Concerns

Feminist economists have also analysed how such factors as race and caste (Brewer et al. 2002) and sexual preference (Badgett 2001), in interaction with gender, affect economic outcomes.

Feminist economists' scepticism about the adequacy of the image of 'economic man' has also stimulated new thinking in areas other than gender relations. The analysis of relations of power and of care, first generated by study of women's work and family relations, has been extended to the subject of interpersonal relations among people economy-wide (Nelson 2005). Feminist explorations into ecological economics examine how natural processes, like women, have been treated as invisible and freely exploitable in traditional economic thought (Perkins 1997).

See Also

- ▶ Economic Man
- ▶ Gender Roles and Division of Labour
- ▶ Household Production and Public Goods
- ▶ Intrahousehold Welfare
- ▶ Methodology of Economics
- ▶ Sen, Amartya (Born 1933)
- ▶ Women's Work and Wages

Bibliography

Aerni, A., and K. McGoldrick (eds.). 1999. *Valuing us all: Towards feminist pedagogy in economics*. Ann Arbor: University of Michigan Press.

Agarwal, B. 1997. 'Bargaining' and gender relations: Within and beyond the household. *Feminist Economics* 3(1): 1–51.

Agarwal, B., J. Humphries, and I. Robeyns. 2003. A special issue on Amartya Sen's work and ideas. *Feminist Economics* 9(2/3).

Albelda, R., S. Himmelweit, and J. Humphries. 2004. A special issue on lone mothers. *Feminist Economics* 10(2).

Aslanbeigui, N., S. Pressman, and G. Summerfield (eds.). 1994. *Women in the age of economic transformation*. London: Routledge.

Badgett, L. 2001. *Money, myths and change: The economic lives of lesbians and gay men*. Chicago: University of Chicago Press.

Balakrishnan, R. (ed.). 2002. *The hidden assembly line: Gender dynamics of subcontracted work in a global economy*. Bloomfield: Kumarian Press.

Barker, D., and E. Kuiper (eds.). 2003. *Toward a feminist philosophy of economics*. London/New York: Routledge.

Benería, L. 2003. *Gender, development, and globalization*. London: Routledge.

Bergmann, B. 1974. Occupational segregation, wages and profits when employers discriminate by race or sex. *Eastern Economic Journal* 1(2/3): 103–110.

Bergmann, B. 1989. Does the market for women's labor need fixing? *Journal of Economic Perspectives* 3(1): 43–60.

Bettio, F., and J. Plantenga. 2004. Comparing care regimes in Europe. *Feminist Economics* 10(1): 85–113.

Booth, A., J. Burton, and K. Mumford. 2000. The position of women in UK academic economics. *Economic Journal* 110: 312–333.

Brewer, R., C. Conrad, and M. King, eds. 2002. A special issue on gender, color, caste and class. *Feminist Economics* 8(2).
Budlender, D., D. Elson, G. Hewitt, and T. Mukhopadhyay. 2002. *Gender budgets make cents*. London: Commonwealth Secretariat Publications.
Çagatay, N., D. Elson, and C. Grown, eds. 1995. Special issue on gender, adjustment and macroeconomics. *World Development* 23(11).
Danby, C. 2004. Toward a gendered Post Keynesianism. *Feminist Economics* 10(3): 55–75.
Dimand, M.A., R.W. Dimand, and E.L. Forget (eds.). 1995. *Women of value: Feminist essays on the history of women in economics*. Aldershot: Edward Elgar.
Elson, D. (ed.). 1991. *Male bias in the development process*. Manchester: Manchester University Press.
Emami, Z. 1993. Challenges facing social economics in the twenty-first century: A feminist perspective. *Review of Social Economy* 52: 416–425.
Feiner, S. 2004. There are none so blind. . .. In *A guide to what's wrong with economics*, ed. E. Fullbrook. London: Anthem Press.
Ferber, M., and B. Birnbaum. 1977. The 'new home economics': Retrospects and prospects. *Journal of Consumer Research* 4: 19–28.
Ferber, M., and J. Nelson (eds.). 1993. *Beyond economic man: Feminist theory and economics*. Chicago: University of Chicago Press.
Ferber, M., and J. Nelson (eds.). 2003. *Feminist economics today: Beyond economic man*. Chicago: University of Chicago Press.
Figart, D., E. Mutari, and M. Power. 2002. *Living wages, equal wages*. London: Routledge.
Folbre, N. 1991. The unproductive housewife: Her evolution in nineteenth-century economic thought. *Signs* 16: 463–485.
Folbre, N. 1994. *Who pays for the kids?* London: Routledge.
Folbre, N., and J. Nelson. 2000. For love or money – Or both? *Journal of Economic Perspectives* 14: 123–140.
Ginther, D., and S. Kahn. 2004. Women in economics: Moving up or falling off the academic career ladder? *Journal of Economic Perspectives* 18(3): 193–214.
Grown, C., D. Elson, and N. Çagatay, eds. 2000. Special issue on growth, trade, finance and gender inequality. *World Development* 28(7).
Hartmann, H. 1976. Capitalism, patriarchy and job segregation by sex. *Signs* 1(3, part 2): 137–169.
Himmelweit, S. 1999. Caring labor. *Annals of the American Academy of Political and Social Science* 561: 27–38.
Humphries, J. 1990. Enclosures, common rights, and women. *Journal of Economic History* 50: 17–42.
Ironmonger, D. 1996. Counting outputs, capital inputs and caring labor. *Feminist Economics* 2(3): 37–64.
Kabeer, N. 2000. *The power to choose: Bangladeshi women and labour market conditions in London and Dhaka*. London: Verso.
Kuiper, E., and J. Sap (eds.). 1995. *Out of the margin: Feminist perspectives on economics*. London: Routledge.
MacDonald, M. 1998. Gender and social security policy. *Feminist Economics* 4(1): 1–25.
Matthaei, J. 1996. Why feminist, marxist and anti-racist economists should be feminist-marxist-anti-racist economists. *Feminist Economics* 2(1): 22–42.
McElroy, M., and M. Horney. 1981. Nash bargained household decisions. *International Economic Review* 22: 333–349.
Nelson, J. 1995. Feminism and economics. *Journal of Economic Perspectives* 9(2): 131–148.
Nelson, J. 2005. Interpersonal relations and economics. In *Economics and social interactions*, ed. B. Gui and R. Sugden. Cambridge: Cambridge University Press.
Perkins, E., ed. 1997. Women, ecology and economics. Special issue of *Ecological Economics* 20(2).
Peterson, J., and M. Lewis. 1999. *The Elgar companion to feminist economics*. Cheltenham: Edward Elgar.
Pujol, M. 1992. *Feminism and anti-feminism in early economic thought*. Aldershot: Edward Elgar.
Rubery, J. (ed.). 1998. *Equal pay in Europe?* London: Macmillan.
Sen, A. 1984. *Resources, values, and development*. Cambridge, MA: Harvard University Press. ch. 16.
Sen, A. 1990. More than 100 million women are missing. *The New York Review of Books*, 20 December.
Shackelford, J. 1992. Feminist pedagogy. *American Economic Review* 82: 570–576.
Strassmann, D. 1995. Editorial: Creating a forum for feminist inquiry. *Feminist Economics* 1(1): 1–5.
Wagman, B., and N. Folbre. 1996. Household services and economic growth in the United States, 1870–1930. *Feminist Economics* 2: 43–66.
Waller, W., and A. Jennings. 1990. On the possibility of a feminist economics. *Journal of Economic Issues* 24: 613–672.
Waring, M. 1988. *If women counted*. San Francisco: Harper & Row.
Wheelock, J., E. Oughton, and S. Baines. 2003. Getting by with a little help from your family: Toward a policy-relevant model of the household. *Feminist Economics* 9(1): 19–45.

Ferguson, Adam (1723–1815)

Nicholas Phillipson

Ferguson was born in Perthshire in 1723 and died in Edinburgh in 1815. He was educated at St Andrew's University for the Church of Scotland and became a leading member of the 'moderate' clergy which controlled its affairs from 1752 to 1805. He was a charismatic teacher who held the

Moral Philosophy chair at Edinburgh from 1764 to 1785, transforming its curriculum and laying the foundations of its international reputation. As a moralist, Ferguson was worried by the materialism inherent in modern philosophy and modern life, and was anxious to show that the classical republicanism of the Machiavellians was still of value in analysing and resolving its problems. He presented human beings as active rather than passive agents who were motivated by a natural love of perfection that seemed to be in danger of extinction in a commercial world. In the process he showed that the mechanics of social bonding in primitive societies in particular were more complex than contemporaries realized, a demonstration that continues to be admired by anthropologists.

Marx admired Ferguson's discussion of the division of labour and the apparent alienation that accompanied its progress and he thought that Smith's treatment of the subject owed much to him. In fact the resemblances are only superficial. Ferguson's treatment of the subject and of political economy generally was derivative and shaped by the classical republican's traditional concern with virtue, corruption and the place of the heroic virtues in an age of commerce. *The Wealth of Nations* was to leave no significant marks on his thought. He was a moralist who sought to tighten, not loosen the ties which bound political economy to moral philosophy.

Ferguson's contemporary reputation rested on three frequently republished and translated works, *An Essay on the History of Civil Society* (1769) and *The History of the Progress and Termination of the Roman Republic* (1783). His lectures were published as *The Principles of Moral and Political Science* (1792).

See Also

► Enlightenment, Scottish

Selected Works

1767. *An essay on the history of civil society,* ed. D. Forbes. Edinburgh: University Press, 1966.

References

Kettler, D. 1965. *The social and political thought of Adam Ferguson*. Columbus: Ohio State University Press.

Ferrara, Francesco (1810–1900)

F. Caffé

Ferrara was not only an economist but also an influential figure in Italian politics, culture and journalism. He was born in Palermo on 7 December 1810 and died in Venice on 22 January 1900. His long life spanned the political unification of Italy and the country's first attempts to assert itself as a latecomer to the international scene. As a patriot, he was one of the leaders of the Sicilian revolution against the Bourbons in 1848. Although the failure of this uprising led to the return of the Bourbons and subjected Ferrara to exile in Turin, one of the most significant documents of this period is the *Letter from Malta*, which constituted a formal indictment of the Bourbon government and is attributed to him. In Turin, Ferrara became a friend of Cavour, and he was appointed Professor of Political Economy at the university there. As soon as Sicily was liberated he returned to Palermo, where he was placed in charge of indirect taxation.

In 1862 he went back to Turin to assist Quintino Sella, the founder of Italian public finance, in the formulation of fundamental laws to resolve, through harsh and unpopular measures, the financial difficulties of the time. For a brief period in 1867 he was Minister of Finance, and he subsequently became a member of the Chamber of Deputies until 1880. During this time of political and parliamentary activity he also produced a prodigious amount of important journalistic work, inspired by his guiding principle that 'economics was the new way of the necessity of freedom'. His intransigent and uncompromising liberalism placed him in direct contrast with Cavour and contributed to his position as a respected but isolated figure.

His economic achievements are such that he is highly regarded by all the greatest Italian economists (Pareto, Pantaleoni, Einaudi, Del Vecchio) as having inspired and created an Italian economic school of thought. Ferrara has a double claim to this. As the founder of an *Economists' Library*, of which he edited the first two series, he brought the greatest foreign economists within the sphere of Italian cultural life, writing perceptive prefaces to the translations of their works. Guided by his extraordinary knowledge of economic thought, he included translations of the works of authors such as Henry Carey at a time when that author's work was forgotten even by his fellow-Americans. As a theoretician, Ferrara pursued vigorous polemics against the German historical school, which rejected the theoretical method in economics. He developed the concept of cost of reproduction based on technical and psychological factors which pre-dated the marginalist theory in all but name. He was also an early forerunner of the Italian tradition in the economic foundations of public finance.

However, Ferrara has yet to achieve his rightful recognition at an international level. Praised by G.H. Bousquet, who edited the French translation of his selected works, criticized by Schumpeter for his 'ultraliberalism', he is unknown to modern proponents of neo-liberalism, even though he was the forerunner of many of their proposals for de-regulation; for example, the free creation of money without state interference, which has been recently advocated by Hayek. Another aspect of Ferrara's thinking that finds an immediate place in today's economic debates is his conception of 'generalized crowding out', which he formulated not only with regard to the predominant absorption of financial flows by the public sector (and therefore to the detriment of private enterprise) but with respect to every form of public intervention.

Selected Works

Part of *Le opere complete di Francesco Ferrara* was published in homage to Luigi Einaudi, on the occasion of his 80th birthday. This edition, which was edited with great philological rigour by Bruno Rossi Ragazzi, who died prematurely, and was continued by various other acedemics, has yet to be completed. In its present state it comprises ten volumes, which include his *Prefazioni alla biblioteca dell' Economista,* his *Articoli su giornali*, his *Saggi rassegne memorie economiche e finanziarie* and *Discorsi parlamentari*. Not included are the *Lezione de Economia Politica*, which were given by him and which are available in a two-volume edition edited by G. De Mauro Tesoro (1934–5), Zanichelli: Bologna. A collection of *Oeuvres économiques choisies* by Ferrara has been edited by G.H. Bousquet and J. Crisafulli (1938), Paris: Rivière. See also G.H. Bousquet (1960), *Esquisse d'une historie de la science économique en Italie – des origines à Francesco Ferrara*, Paris: Rivière.

Fertility

Richard A. Easterlin

At the aggregate level, human reproduction is the ultimate source of an economic system's labour input and of the consumers who constitute the principal destination of the economy's output. At the individual level, children are an important source of satisfaction that compete with alternatives for the limited parental resources of time, energy and money available. Despite this, reproductive behaviour has traditionally been omitted from economic theorizing, and even in the past three decades has gained only a marginal foothold.

Possibly the hesitancy of economic theory to address the determinants of childbearing reflects a sensitivity to reality. Several empirical regularities involving the relation of fertility to income have posed a formidable challenge to theoretical interpretation. First, there is the long-term trend. In an historical epoch when real income per capita and,

in consequence, real consumption of almost all goods has risen at unprecedented rates for a century or more in developed countries, births per couple over the reproductive career have fallen from levels often as high as six or more to two or less. Second, the cross-sectional relation between fertility and income within countries has been found to be variable and often lacks any significant association. Third, over the business cycle a positive association between fertility and income has typically been observed. Moreover, in a number of developed countries in the post-World War II period there was an unprecedented and unanticipated 'baby boom' of a decade or more in duration followed by an almost equally startling 'baby bust'.

It is often the case that new policy concerns stimulate economic theory, and this is clearly so with regard to fertility behaviour. Although unusually low fertility in the developed countries in the Great Depression had led to some experimentation with economic incentives to childbearing, the major policy stimulus by far was the emergence in the post-World War II era of the so-called 'population problem' as a presumed obstacle to economic growth in the less-developed world. Could measures be designed to lower reproduction rates in high-fertility societies and thus reduce rates of population growth? Concern with this issue spurred a number of economists to take a fresh look at fertility behaviour.

The contemporary economic theory of fertility dates from work by Harvey Leibenstein (1957) and Gary S. Becker (1960), which sought in somewhat different ways to assimilate the explanation of fertility behaviour to the economic theory of household demand (Leibenstein (1974) and Keeley (1975) give a good review of this early history). In 1965 Becker extended his analysis to incorporate the emerging concepts of household production theory and the allocation of time (Becker 1965; Lancaster 1971). For a decade or so this line of work, which came to be known as the Chicago-Columbia approach, dominated the economic theory of fertility. Among the more influential contributions were those made by Mincer (1963), Nerlove (1974) and Willis (1973). A volume edited by T.W. Schultz (1974) and a survey article and subsequent book by T.P. Schultz (1976, 1981) brought together a number of attempts to apply this approach empirically to the experience of developed and less-developed countries. A valuable commentary on the evolution of this work appears in Ben-Porath (1982).

Throughout much of this period, a second line of work was in progress that came to be dubbed the 'Pennsylvania' model (Sanderson 1976, 1980; Behrman and Wolfe 1984). Although this work largely accepted the Chicago-Columbia view as far as it went, it sought to broaden the model to include theoretical and empirical considerations that figured prominently in the sociological and demographic literature on fertility. One set of considerations related to taste influences on the demand for children, particularly of what economists would term a 'relative income' nature (Ben-Porath 1975; Easterlin 1969; Leibenstein 1975). Good theoretical expositions of the demand for children, reflecting taste considerations as well as those in the Chicago-Columbia model, are given by Lindert (1978) and Turchi (1975). The other set of considerations relates to 'natural fertility' or so-called 'supply' factors (Easterlin 1978; Tabarrah 1971). A formal statement of what came to be termed the 'supply-demand' approach was published in 1980 (Easterlin et al. 1980). In 1983 an interdisciplinary National Academy of Sciences panel adopted the supply-demand framework in surveying the literature on determinants of fertility in developing countries (Bulatao and Lee 1983). (As shall become clear, in this theory supply and demand are not used in the usual economic sense.)

Recent work points to some convergence of the two lines of work. Scholars working in the Chicago-Columbia tradition have introduced into their work intergenerational influences which can be likened to the taste influences of the Pennsylvania model (Becker and Tomes 1976), and have also started investigating supply factors (Michael and Willis 1976; Rosenzweig and Schultz 1985). But the two schools remain sufficiently distinct, especially in their interpretation of the empirical regularities described above, to warrant separate discussion. Using the empirical regularities as the framework for the discussion the following aims to

indicate for the non-specialist the principal ideas on each side, their relations to each other, and their bearing on the interpretation of each of the empirical regularities mentioned. Needless to say, there is also variability within each school as well as work not easily classified under either head.

The Secular Decline in Fertility

Over the long term, growth in real per capita income has everywhere been accompanied by a decline in child-bearing from levels sometimes averaging as high as six or more births per woman to around two or less. In seeking to explain this development as in understanding fertility behaviour more generally, the Chicago-Columbia model focuses on changes in the demand for children.

A simple economic model analogous to that for the demand for any economic good, the original starting point for economic theorizing on fertility, would see the number of children demanded as varying directly with household income (assuming children are a 'normal' good), directly with the price of goods relative to children, and inversely with the strength of tastes for goods relative to children. In the Chicago-Columbia approach price and income are the explanatory variables featured, and especially price. The explanation of the secular fertility decline provides a typical illustration. The decline is seen as due to a decrease in the demand for children brought about by socio-economic development. This decrease in demand, in turn, is ascribed to a strong negative effect associated with an increase in the relative price of children that outweighs a weak positive effect from higher income. The most common explanation of the increased relative price of children focuses on the opportunity cost of the wife's time. In keeping with the household production function concept, children are seen as requiring inputs of goods and time, and the price of children, as depending, accordingly, on the prices of these inputs. Typically, the input of the wife's time in childbearing and raising is the central focus, and the assumption is made that children are more time-intensive with regard to the wife's time than other forms of consumption. The opportunity cost of the time input into children is then seen as increasing secularly as the wife's opportunity cost, proxied by her schooling, rises (Willis 1973). Other factors increasing the relative price of children are sometimes cited such as the price of labour relative to capital, the prices of other child inputs, or a systematic change in the 'quality' of children demanded, where quality is identified with the quantities of inputs of time and goods into a child (Lindert 1980, 1983; Schultz 1974, 1979). All of these are seen as working via a negative impact on the demand for children, as in the case of the opportunity cost of a wife's time.

Several objections based on empirical studies have been raised to a purely demand interpretation of the secular fertility decline. For one thing, a phase of increasing fertility has often preceded the secular fertility decline (Dyson and Murphy 1985); is this to be taken as implying an initial period of increasing demand for children? Then, too, there are indications of low fertility in some sub-Saharan African societies, apparently associated with venereal disease. How is this to be treated in a demand-oriented theory? Perhaps most important, demographic surveys of contemporary premodern populations have repeatedly turned up evidence of what demographers call a 'natural fertility' regime, the absence of any attempt deliberately to limit fertility among almost all segments of the population, aside from some elite groups (Coale 1967; Henry 1961). If parents are choosing the number of children they have in accordance with a demand model, how is one to explain the fact that so few are doing anything to control their fertility? It seems unlikely that unregulated fertility would assure that most couples would have just as many children as they want and no more, and thus result almost uniformly throughout a population in no practice of family size limitation.

To deal with questions of this type the Pennsylvania model stresses two factors as fertility determinants in addition to the demand for children: (1) the potential supply of children, the number of surviving children parents would have if they did not deliberately limit fertility; and (2) the costs of fertility regulation, including

both subjective (psychic) drawbacks and objective costs, the time and money required to learn about and use specific techniques.

The introduction of supply considerations is the most distinctive feature of the Pennsylvania model. (Regulation costs are sometimes treated in Chicago-Columbia models although usually subordinated to demand.) The most obvious example of the importance of a supply constraint in determining observed fertility is the case of a couple that has fecundity problems and is consequently unable to produce as many children as it wants. True, child adoption is a logical option in such a case, but empirically this practice is of quite limited significance. Clearly, a supply constraint due to sterility would explain the African case mentioned above. Aside from sterility problems, however, the production of children is kept down significantly in almost all pre-modern societies by various types of behaviour that have unperceived consequences for fertility (a phenomenon conceptually designated 'unperceived jointness' by Easterlin et al. 1980). The most important types of behaviour in this regard are deferment of sexual unions beyond menarche, which reduces exposure to intercourse and prolonged breast feeding, which has the effect of delaying the return of ovulation after a birth. Also, because the subject of parents' demands is not births per se, but surviving children, high infant and child mortality in pre-modern societies further restricts the supply of children.

The Pennsylvania model suggests two possible reasons for 'natural fertility' behaviour. First, there is the possibility of excess demand. If in most households in a pre-modern society the supply of children were less than demand, then parents would have as many children as they could. In such a situation there would be a general absence of any practice of deliberate family size limitation, and differences in observed fertility would be determined by the circumstances responsible for differences in supply.

Even if supply exceeded demand, however, deliberate family size limitation would not necessarily occur, because of the costs attaching to the various techniques of fertility control. If, for example, the disutility attaching, say, to abstinence or withdrawal, exceeds the disutility of an excess number of children, and no other contraceptive practices are known, then a couple's observed fertility would again be governed by its supply. Thus the Pennsylvania model identifies two cases in which rational behaviour would be consistent with an absence of deliberate fertility regulation – (a) an excess demand condition, and (b) high perceived costs of fertility regulation.

In interpreting the secular fertility decline, the Pennsylvania model envisages a typical pre-modern society as starting from a condition of unregulated fertility due to either or both of the circumstances just mentioned, and moving to a situation of algebraically increasing excess supply, as supply increases and demand decreases (though not necessarily concurrently) with socio-economic development. The increase in supply might reflect a decrease in breastfeeding that raises natural fertility, improved child survival, or both. The decreased demand might be due to a rise in the relative cost of children, as in the Chicago–Columbia model, or an anti-natal shift in tastes due to education or the introduction of new consumer goods (Behrman and Wolfe 1984; Easterlin 1978). If the disutility associated with an excess supply of children remains less than the disutility associated with use of contraception, then fertility will remain uncontrolled and an increase in actual fertility, reflecting the growth of natural fertility, will be observed. Such circumstances could account for a phase of increasing fertility prior to the secular fertility decline. Eventually, however, as excess supply continues to rise, the pressures for adoption of deliberate control will prevail, and observed fertility will decline. Lower costs of fertility regulation due, say, to better contraceptive knowledge or improved contraceptive availability, might also contribute to the shift to lower fertility, although it is unlikely that this factor in itself could explain an initial phase of increasing fertility.

Thus, the supply-demand approach sees the supply of children and regulation costs, as well as demand, as factors that may significantly influence observed fertility in pre-modern societies and in the early stages of the secular fertility decline. Eventually, however, as modernization

progresses, most households shift to a position of substantial potential excess supply and increasingly perceive costs of fertility regulation as low. Because of this, demand influences become increasingly dominant in determining fertility, although these influences may reflect taste changes in addition to anti-natal price effects (cf. Mueller and Short 1983). Thus, in contemporary developed societies, both schools emphasize demand as the principal influence determining fertility, although they differ with regard to the underlying determinants of demand that are considered most important.

In recent work by some members of the Chicago-Columbia school, a supply-demand model has also been adopted, but sharp conceptual differences from the Pennsylvania model remain (Schultz 1981; Rosenzweig and Schultz 1985). In the Chicago–Columbia model supply is determined solely by biological factors and all behavioural factors operate through demand. In contrast, in the Pennsylvania model, supply is constrained by behaviour that has the unintentional effect of limiting fertility. Particularly at issue is the extent to which considerations of family size enter into individual decisions on marriage-timing, length of breastfeeding, and consumption. Based on motivational data from sample surveys as well as behavioural data of various types, the Pennsylvania model assumes such decisions to be taken largely independently of family size concerns. The positivist methodology of the Chicago–Columbia school leads to rejection of this evidence, and to a theoretical conception that stresses on a priori grounds the endogeneity to the fertility decision of marriage-timing and breastfeeding behaviour. In the Pennsylvania model it is the behavioural influences on supply that are principally responsible for constraining fertility to the extent that it may fall short of demand in premodern societies. Aside from cases of physiological sterility, the supply concept of the Chicago–Columbia school would be unlikely to constrain fertility effectively; hence their supply-demand model largely preserves their emphasis on demand conditions as the determinant of observed fertility behaviour in all times and places.

The Cross-Sectional Relation Between Fertility and Income

As has been mentioned, cross-sectional empirical studies of the relation between fertility and income yield mixed results, and this is so even after controlling for numerous other variables (Mueller and Short 1983; Simon 1974). Sometimes the direction of relationship is positive, sometimes negative; sometimes the relationship is significant, sometimes not significant. Because economists of both schools working in the fertility area intuitively accept that children are a normal good, these empirical findings have provoked an extensive research for explanation, and, in particular, for price or taste factors that might vary systematically with income level. This search has again involved rather different types of emphasis by the Chicago-Columbia and Pennsylvania schools.

Consistent with the interpretation of the secular fertility decline, the variable stressed most frequently by the Chicago-Columbia school has been the opportunity cost of a wife's time, a variable first brought to the fore by Mincer 1963). The idea is that husbands with higher permanent income are likely to have spouses with higher education and thus higher market wage rates. A general increase in the potential earnings of both sexes would then lead to a substitution effect against children due to the wife's higher wage rate that offsets a positive income effect from the husband's wage rate. Because of the absence of wage rates for nonworking wives, empirical tests of this hypothesis have usually used wife's education as a proxy for the opportunity cost of wife's time (Schultz 1974).

A second line of explanation in the Chicago-Columbia approach stressed by Becker himself and differing from the price-of-time argument emphasizes the association between what is called 'child quality' and income (Becker and Lewis 1974). In this case the variation with income of the quantities of child inputs, rather than their prices, is the focus. The basic idea is that as parents' income increases, they are assumed to want to increase child inputs and thus to spend more, on the average, on their children, just as

they are expected to want to spend more on themselves. This positive association between desired expenditures per child and parental income causes children to be more expensive for wealthier parents than poorer ones, and is presumed to offset the positive effect of income per se on the demand for children.

Although the Pennsylvania school does not reject the plausibility of these hypotheses, it offers yet another one, again influenced by the demographic work of sociologists, as in the supply-demand approach. In this case, the analysis builds on the sociological notion that one's economic socialization experience early in the life cycle plays an important part in forming one's material tastes. It is assumed that the material environment surrounding young persons in the course of their upbringing leads to the formation of a socially defined subsistence level that they wish to achieve on reaching adulthood (Ahlburg 1984; Easterlin 1973). Only to the extent that actual income exceeds this subsistence constraint would a couple feel free to embark on family formation. Assuming that young adults with higher income come from more affluent backgrounds, then the expected positive effect of higher income on the demand for children would be offset by the higher goods aspirations of wealthier compared with poorer couples (Easterlin 1969).

Recently there has been some convergence in the two views. On the one hand, the Chicago-Columbia school has introduced consideration of what is called 'child endowments', stressing the effect of one's family of origin on fertility behaviour, as in the Pennsylvania model (Becker 1981; Becker and Tomes 1976; Nerlove 1974). On the other hand, the Pennsylvania school has added to its conception concern with parents' desires for expenditures on their children as well as for themselves (Easterlin 1976). Both schools typically argue that because of such systematic correlates of changing income, the cross-sectional observed relation of income and fertility is uncertain. Leibenstein (1974, 1975) takes a stronger stance, asserting on a priori grounds that negative taste changes associated with higher income offset positive income effects and cause a negative association between income and fertility. The argument is that a higher status household must more than proportionately increase its expenditures on 'status goods' in order to maintain its relative life style and economic status.

Much of the theorizing regarding the cross-sectional income/fertility relationship assumes that the same arguments would apply in both pre-modern and developed societies. Development of the supply–demand model by the Pennsylvania school has led to reconsideration of this proposition. If, in pre-modern societies, fertility is largely determined by supply conditions, whereas in developed countries it is largely determined by demand factors, there is obviously no reason to expect the same relation between income and fertility (Behrman and Wolfe 1984). Indeed, a systematic shift in the relation of income to fertility might be observed, as the dominant determinants shifted from supply to demand (Crimmins et al. 1984). To illustrate, from a supply viewpoint, in a pre-modern society higher income might lead to better nutrition and thereby higher fecundity of a wife, or to shorter breastfeeding as baby food substitutes become available and affordable. For both reasons the ability to produce children would be positively related to income. If a natural fertility regime prevailed, then such supply effects might yield a positive relation between observed fertility and income. With the progress of modernization and a growing predominance of demand factors in fertility determination, this initial positive cross-sectional association might change to a non-significant or (on Leibenstein reasoning) even a negative relationship. The possibility that supply factors might dominate the cross-sectional income-fertility relationship in developing countries has not been considered by the Chicago-Columbia school, presumably because of their more restricted concept of supply.

Fluctuations

Economic theorizing about fertility fluctuations has focused primarily on the experience of the developed countries, and particularly the United States. The protracted postwar baby boom and bust, from a low in the 1930s to a peak in the

1950s and then a new 1970s trough, has attracted most attention; shorter-term business cycle fluctuations, much less.

The interpretation of the United States baby boom and bust advanced by the Pennsylvania school is a relative income one that builds on the arguments about taste formation described in the preceding section (Easterlin 1973, 1980). The basic idea is that the cohorts that were in the family forming ages in the late 1940s and 1950s were raised under the economically deprived circumstances of the Great Depression and World War II. As a result, the material aspirations formed during their economic socialization experience were low. Their labour market experience, however, was quite favourable, because of the combined circumstances of a prolonged post-World War II economic expansion and the relative scarcity of young workers, the latter echoing the unprecedentedly low fertility of the 1920s and 1930s. In consequence, these cohorts enjoyed high relative income, that is, high income relative to their material aspirations, and this led to earlier marriage and child-bearing, higher completed family size, and the baby boom that lasted through the late 1950s.

The circumstances of the subsequent cohorts tended to be the reverse – declining relative income, postponed marriage and childbearing, lower completed family size – adding up to a baby bust. On the one hand, these cohorts had formed high material aspirations as a result of their upbringing in the boom circumstances following World War II. On the other, their own labour market experience was much less favourable, partly because of some slackening in the growth of aggregate demand, and partly because of a sharply increased relative supply of workers in family forming ages, itself a consequence of the prior baby boom.

As the foregoing suggests, the relative income hypothesis can with some restrictive assumptions, be translated into a relative cohort size hypothesis. If one assumes fairly stable growth in aggregate demand and a largely closed economy, then variations in the earnings of younger compared with older workers would be dominated by variations in the relative supply of younger workers. A relatively small cohort of young adults would cause a narrowing of the shortfall of younger workers' incomes compared with older; a relatively large cohort would cause a widening of the gap. Taking older workers' incomes as a proxy for the material aspirations formed by young adults when in their parents' homes, one obtains the same type of relative income mechanism engendering fertility movements that was just described. This relative cohort size variant has been used to demonstrate the possibility of a self-generating fertility cycle (Lee 1974; Samuelson 1976).

In contrast to the taste formation influences emphasized in the Pennsylvania model, the Chicago-Columbia interpretation of the baby boom and bust builds on a price-of-time argument similar to that used in explaining the secular fertility decline (Butz and Ward 1979). An increase in husband's income is thought to have a positive effect on fertility; an increase in the wife's wage rate, a negative effect due to the price-of-time effect. In the baby boom period, it is argued, the labour market for women relative to men, as indexed by wage rate movements for the two sexes, was comparatively weak; thereafter the labour market for women expanded commensurately with that for men. Thus, in the baby boom period, men's wage rates rose while women's remained relatively flat; hence a net positive impact on fertility prevailed, reflecting the dominant effect of men's compared with women's wage rate changes. Thereafter, women's wage rates rose commensurately with men's, and a negative effect dominated, due to the higher absolute magnitude of the elasticity of fertility with respect to women's wage rates than men's. The result of the disparate changes in men's and women's wages before and after 1960 was thus an upswing in fertility followed by a downswing. Young women's labour force participation moved inversely with fertility in the two periods, reflecting the differing pull of women's wage rates.

Several critiques of the econometric techniques used in this analysis have appeared (Kramer and Neusser 1984; McDonald 1983). Also, the movement in labour force participation of older women does not fit easily into the argument. This is because older women, who are

highly substitutable for younger women in most jobs, showed a marked rise in labour force participation before 1960, the period when the female labour market was presumably weak, and then much slower growth after 1960, when the female labour market was presumably stronger. In addition, the Chicago–Columbia view implies that the more favourable movement in women's wages after 1960 would have shortened birth intervals (Mincer and Polachek 1974), whereas the opposite, in fact, occurred. Nevertheless, the Butz–Ward analysis remains the prevailing interpretation advanced by adherents of the Chicago–Columbia school. Some analysts have found that a combination of the Pennsylvania and Chicago-Columbia models is superior to either alone (Devaney 1984; Lindert 1978).

The Pennsylvania and Chicago–Columbia models have quite different implications for the future of fertility fluctuations. The Pennsylvania model suggests that a growing scarcity of younger workers in the 1980s and 1990s, echoing the baby bust of the 1960s and 1970s, is a factor making for a turnaround in the relative income of young adults and thus for an upturn in fertility. In contrast, the Chicago–Columbia view envisages further fertility declines on the assumption that women's wages are likely to continue to rise commensurately with men's.

The two models also differ in their predictions regarding variations in fertility over the business cycle. The Pennsylvania model anticipates a positive association of the type traditionally observed (Ben-Porath 1973). Because of the importance of historical experience in forming tastes, one would expect tastes to remain largely invariant in periods as short as the usual business cycle. Variations in actual earnings associated with the business cycle might therefore be expected to lead to corresponding variations in fertility as income expectations were revised.

In contrast, the Chicago–Columbia analysis suggests that the short-term association between fertility and income varies with the proportion of females in the labour force. The reasoning is that the relative importance of the negative effects of women's wage changes versus the positive effect of men's wage changes will be greater, the larger the proportion of women in the labour force. Hence, as the proportion of women at work rises, the more sensitive does fertility become to fluctuations in women's wage rates. Based on the uptrend in women's labour force participation, the Chicago-Columbia model thus foresees the emergence of counter-cyclical fertility fluctuations. When women's wage rates are high, as in a boom period, the price-of-time effect will pull them into the labour market and consequently reduce fertility; when wage rates are low, the reverse will occur. This mechanism is claimed to have operated during the business cycles of the 1970s.

Conclusion

This survey has aimed at highlighting some of the principal differences between the two leading schools of economic theorizing about reproductive behaviour, as manifested in the interpretations offered of trends, fluctuations and cross-sectional variations in the income-fertility relationship. As the survey demonstrates, the evolution of theorizing on reproductive behaviour has been away from a simple economic model of demand emphasizing income and market price variables toward the recognition of additional constraints on behaviour. Perhaps more than in any other area of economic analysis, the constraint of time inputs has come to the forefront, both the amounts of time required in childbearing and childrearing and the prices at which these inputs should be valued. This interest has stimulated fruitful empirical inquiries by economists into the use of time within the household. Also, explicit attention has been paid to the constraint on one's behaviour arising from the way that prior experience shapes one's tastes. Thus, attention has been directed to the way that one's experience in one's family of origin and one's socialization experience more generally may shape adult preferences with regard to material aspirations and family size. In this area, economists have sometimes pushed beyond the conceptual speculations of sociologists to formulate specific empirical models of taste formation. Finally, recognition has emerged of the constraint on 'consumption' of children arising from

production possibilities within the household. Whether because of biological or behavioural attributes, a couple may be unable to produce as many children as demanded, and its observed consumption would thus reflect this rationing constraint.

As the foregoing discussion shows, the introduction of these new behavioural constraints has arisen from growing awareness by economists of the intractability of empirical evidence on reproductive behaviour, and a resultant attempt to accommodate within economic analysis conceptual contributions from related disciplines. Although some progress in understanding reproductive behaviour has been made, there is still no single generally accepted theory of reproductive behaviour, and no consensus on the interpretation of the empirical regularities described above. However, if it is true that scientific breakthroughs frequently occur at the juncture of different disciplines, fertility theory is undoubtedly one of the frontiers of economic theory beckoning for more intensive exploration.

See Also

- ▶ Demographic Transition
- ▶ Demography
- ▶ Easterlin Hypothesis
- ▶ Family Planning
- ▶ Fecundity
- ▶ Historical Demography
- ▶ Infant Mortality
- ▶ Stable Population Theory

Bibliography

Ahlburg, D.A. 1984. Commodity aspirations in Easterlin's relative income theory of fertility. *Social Biology* 31(3/4): 201–207.

Becker, G.S. 1960. An economic analysis of fertility. In *Demographic and economic change in developed countries.* Universities-National Bureau conference series No. 11. Princeton: Princeton University Press.

Becker, G.S. 1965. A theory of the allocation of time. *Economic Journal* 75: 493–517.

Becker, G.S. 1981. *A treatise on the family.* Cambridge, MA: Harvard University Press.

Becker, G.S., and H.G. Lewis. 1974. Interaction between quantity and quality of children. In *The economics of the family*, ed. T.W. Schultz. Chicago: University of Chicago Press.

Becker, G.S., and N. Tomes. 1976. Child endowments and the quantity and quality of children. *Journal of Political Economy* 84(4), Part 2: S143–S162.

Behrman, J.R., and B.L. Wolfe. 1984. A more general approach to fertility determination in a developing country: The importance of biological supply considerations, endogeneous tastes and unperceived jointness. *Economica* 51: 319–339.

Ben-Porath, Y. 1973. Short-term fluctations in fertility and economic activity in Israel. *Demography* 10(2): 185–204.

Ben-Porath, Y. 1975. First generation effects on second generation fertility. *Demography* 12(3): 397–405.

Ben-Porath, Y. 1982. Economics and the family-match or mismatch? A review of Becker's 'a treatise on the family'. *Journal of Economic Literature* 20(1): 52–64.

Bulatao, R.A., and R.D. Lee (eds.). 1983. *Determinants of fertility in developing countries: A summary of knowledge*. New York: Academic Press.

Butz, W.P., and M.P. Ward. 1979. The emergence of countercyclical US fertility. *American Economic Review* 69(3): 318–328.

Coale, A.J. 1967. The voluntary control of human fertility. *Proceedings of the American Philosophical Society* 111(3): 164–169.

Crimmins, E.M., R.A. Easterlin, S.J. Jejeebhoy, and K. Srinivasan. 1984. New perspectives on the demographic transition: A theoretical and empirical analysis of an Indian state, 1951–1975. *Economic Development and Cultural Change* 32(2): 227–253.

Devaney, B. 1984. An analysis of variations in U.S. fertility and female labor force participation trends. *Demography* 20(2): 147–161.

Dyson, T., and M. Murphy. 1985. The onset of fertility transition. *Population and Development Review* 11(3): 399–440.

Easterlin, R.A. 1969. Towards a socioeconomic theory of fertility: A survey of recent research on economic factors in American fertility. In *Fertility and family planning: A world view*, ed. S.J. Behrman, Leslie Corsa Jr., and R. Freedman. Ann Arbor: University of Michigan Press.

Easterlin, R.A. 1973. Relative economic status and the American fertility swing. In *Family economic behavior: Problems and prospects*, ed. E.B. Sheldon. Philadelphia: Lippincott.

Easterlin, R.A. 1976. Population change and farm settlement in the northern United States. *Journal of Economic History* 36(1): 45–75.

Easterlin, R.A. 1978. The economics and sociology of fertility: A synthesis. In *Historical studies of changing fertility*, ed. C. Tilly. Princeton: Princeton University Press.

Easterlin, R.A. 1980. *Birth and fortune*. New York: Basic Books.

F

Easterlin, R.A., R.A. Pollak, and M.L. Wachter. 1980. Toward a more general economic model of fertility determination: Endogenous preferences and natural fertility. In *Population and economic change in developing countries*, ed. R.A. Easterlin. Chicago: University of Chicago Press.

Henry, L. 1961. La fécondité naturelle: observations – théorie – résultats. *Population* 16(4): 625–636.

Keeley, M. 1975. A comment on 'An interpretation of the economic theory of fertility'. *Journal of Economic Literature* 13(2): 461–467.

Kramer, W., and K. Neusser. 1984. The emergence of countercyclical U.S. fertility: Note. *American Economic Review* 74(1): 201–202.

Lancaster, K.J. 1971. *Consumer demand: A new approach*. New York: Columbia University Press.

Lee, R. 1974. The formal dynamics of controlled populations and the echo, the boom and the bust. *Demography* 11(4): 563–585.

Leibenstein, H. 1957. *Economic backwardness and economic growth*. New York: Wiley.

Leibenstein, H. 1974. An interpretation of the economic theory of fertility: Promising path or blind alley? *Journal of Economic Literature* 12(2): 457–479.

Leibenstein, H. 1975. The economic theory of fertility decline. *Quarterly Journal of Economics* 89(1): 1–31.

Lindert, P.H. 1978. *Fertility and scarcity in America*. Princeton: Princeton University Press.

Lindert, P.H. 1980. Child costs and economic development. In *Population and economic change in developing countries*, ed. R.A. Easterlin. Chicago: University of Chicago Press.

Lindert, P.H. 1983. The changing economic costs and benefits of having children. In *Determinants of fertility in developing countries: A summary of knowledge*, vol. 1, ed. R. Bulatao and R.D. Lee. New York: Academic.

McDonald, J. 1983. The emergence of countercyclical US fertility: A reassessment of the evidence. *Journal of Macroeconomics* 5(4): 421–436.

Michael, R.T., and R.J. Willis. 1976. Contraception and fertility: Household production under uncertainty. In Conference on Research in Income and Wealth, *Household Production and Consumption.* New York: National Bureau of Economic Research.

Mincer, J. 1963. Market prices, opportunity costs, and income effects. In *Measurement in economics*, ed. C. Christ et al. Stanford: Stanford University Press.

Mincer, J., and S. Polachek. 1974. Family investments in human capital: Earnings of women. In *The economics of the family*, ed. T.W. Schultz. Chicago: University of Chicago Press.

Mueller, E., and K. Short. 1983. Effects of income and wealth on the demand for children. In *Determinants of fertility in developing countries: A summary of knowledge*, vol. 1, ed. R. Bulatao and R.D. Lee. New York: Academic Press.

Nerlove, M. 1974. Household and economy: Toward a new theory of population and economic growth. *Journal of Political Economy* 82(2), Part II: S200–S218.

Rosenzweig, M.R., and T.P. Schultz. 1985. The demand for and supply of births: Fertility and its life cycle consequences. *American Economic Review* 75(5): 992–1015.

Samuelson, P.A. 1976. An economist's non-linear model of self-generated fertility waves. *Population Studies* 30(2): 243–247.

Sanderson, W.C. 1976. On two schools of the economics of fertility. *Population and Development Review* 2(3–4): 469–477.

Sanderson, W.C. 1980. Comment. In *Population and economic change in developing countries*, ed. R.A. Easterlin. Chicago: University of Chicago Press.

Schultz, T. 1976. Determinants of fertility: A microeconomic model of choice. In *Economic factors in population growth*, ed. A.J. Coale. New York: Halsted Press.

Schultz, T.P. 1979. Current developments in the economics of fertility. In *International Union for the Scientific Study of Population: Economic and demographic change: Issues for the 1980's*. Proceedings of the Conference, Helsinki 1978, vol. 3, 27–38. Liège: IUSSP.

Schultz, T.P. 1981. *Economics of population*. Reading: Addison-Wesley Co.

Schultz, T.W. (ed.). 1974. *The economics of the family*. Chicago: University of Chicago Press.

Simon, J.L. 1974. *The effects of income on fertility*. Chapel Hill: University of North Carolina Press.

Tabarrah, R.B. 1971. Toward a theory of demographic development. *Economic Development and Cultural Change* 19(2): 257–277.

Turchi, B.A. 1975. *The demand for children: The economics of fertility in the United States*. Cambridge, MA: Ballinger.

Willis, R.J. 1973. A new approach to the economic theory of fertility behavior. *Journal of Political Economy* 81(2), Part II: S14–S64.

Fertility in Developed Countries

Alicia Adsera

Abstract

After completing the first demographic transition, developed countries experienced a fertility boom in the post-Second World War period. However, after the 1960s fertility rates fell dramatically and now, in 2007, stand below the replacement level of 2.1 births per woman in most of these countries. The entry of women into the workforce, economic development and changes in values and secularization are the causes of this demographic transformation.

Keywords

Capital intensity; Child care; Demographic transition; Family planning; Fertility in developed countries; Household production; Infant mortality; Labour market institutions; Labour supply; Second demographic transition; Unemployment; Wage differentials; Welfare state; Women's work and wages

JEL Classification

A1; B5

Decrease in Fertility Rates

Fertility rates in the developed world started to decline sharply at the end of the nineteenth century. In Europe fertility declined by about 40% between 1870 and 1930, and the majority of Western countries experienced this transition before the Second World War (Lee 2003). The Second World War was followed by a period of increases in fertility. After peaking during the 'baby boom' years of the late 1950s, the average total fertility rate in developed countries fell from an average of 2.8 births per woman in 1960 to 2.0 in 1975 and then to 1.6 in the late 1990s (and below 1.3 in southern Europe). The total fertility rate (TFR) estimates the number of children a woman would bear if she went through her childbearing years exposed to the current age-specific birth rates for women between the ages of 15–44 years. Table 1 presents the evolution of the total fertility rate since 1965–2004 in the most developed economies. In 1965, fertility rates were almost three children in many of these countries and even higher in Canada, Portugal, Iceland and Ireland. During the next decade, rates fell sharply in the richest countries, but they remained above replacement level in southern Europe, Ireland and Iceland. The transition to lower fertility has only occurred in southern Europe since the early 1980s but its speed and extent went beyond what previous countries had experienced. By the mid-1990s fertility rates in these countries were under 1.3, a threshold level used by some demographers to define 'lowest-low fertility' (Kohler et al. 2002). With the exception of the United States, Iceland and Ireland, all advanced countries now have fertility rates well below the replacement rate of 2.1 (the fertility rate needed to sustain a steady level of population) though cross-national differences have remained significant.

Fertility in Developed Countries, Table 1 Total fertility rate in developed countries 1965–2004

	1965	1975	1985	1995	2004
Australia	2.98	2.22	1.89	1.82	1.77
Austria	2.7	1.83	1.47	1.4	1.42
Belgium	2.62	1.74	1.51	1.56	1.64
Canada	3.15	1.85	1.61	1.62	1.53
Denmark	2.61	1.92	1.45	1.81	1.78
Finland	2.47	1.69	1.64	1.81	1.8
France	2.83	1.93	1.81	1.7	1.92
Germany	2.51	**1.45**	**1.28**	1.34	1.37
Greece	2.32	2.32	1.67	1.32	1.31
Iceland	3.71	2.65	1.93	**2.08**	2.03
Ireland	**4.03**	**3.4**	**2.5**	1.85	1.99
Italy	2.59	2.17	1.42	**1.17**	1.33
Japan	**2.14**	1.91	1.76	1.42	**1.29**
Netherlands	3.04	1.66	1.51	1.53	1.73
Norway	2.95	1.98	1.68	1.87	1.81
Portugal	3.15	2.63	1.72	1.4	1.4
Spain	2.97	2.79	1.64	1.18	1.33
Sweden	2.42	1.77	1.74	1.73	1.75
Switzerland	2.61	1.61	1.52	1.48	1.42
UK	2.86	1.81	1.79	1.71	1.77
USA	2.88	1.77	1.84	2.02	**2.05**
Average	2.84	2.05	1.68	1.61	1.64

Note: Maximum and minimum rates in *bold*
Sources: National official statistics and United Nations Population Division, various years

Synthetic indices such as total fertility rates may not provide a precise picture of fertility changes in periods when the younger cohorts of women shift the timing of their fertility to older ages. Both the age at first marriage and the age at first birth have been rising since the early 1980s across the developed world. Delayed maternity may artificially deflate total fertility rate since a larger proportion of births is bound to occur among older mothers over time, but this is not still reflected in the behaviour of women currently in their thirties. Adjustments for these 'tempo

effects' suggest that, even though important for some countries, such as France or the Netherlands in the mid-1980s, they account for only part of the reduction in TFR (Bongaarts 2001). Completed fertility by cohort provides an alternative and more accurate way to measure fertility changes. It computes the mean number of children born to women of a given generation at the end of their childbearing years. Recent data in completed fertility show for most countries a downward trend similar to that in TFR though with more moderate inter-country differences. Among women born in 1965, for example, whereas the Irish are projected to bear around 2.2 children, Spaniards, German and Italians are expected to bear only 1.5 children.

In any case, delayed childbearing itself is likely to imply lower completed fertility. Women who become mothers at a later age are expected to bear fewer children by the end of their fertile life because of both time and fecundity constraints (Kohler et al. 2002). Still, the negative relationship between postponement and completed fertility seems to have weakened somewhat, possibly due to the improvement in reproductive techniques. Even if lifetime childlessness has risen steadily among women born between the 1940s and 1970s, particularly in German-speaking countries, demographers project it will stabilize around 15–20% in these countries.

Quantity and Quality of Children

The basic microeconomic model of fertility (Willis 1973; Becker 1991) identifies a broad set of factors that influence fertility: household preferences over the number and quality of children, their labour supply decisions and their access to family planning. Each one of these factors has been relevant to the dramatic reduction in fertility rates across developed countries as income kept rising during the twentieth century. In addition, with modernization, infant mortality decreased sharply. For parents interested in a certain number of surviving children, the increase in the likelihood of survival of any child born constituted an independent cause of the reduction in fertility.

The quantity–quality model developed by Becker and Lewis (1973) and Willis (1973) provided the first explanation of the observation that the number of children per family did not increase with income. In this model, each family maximizes a utility function depending on the quantity of children, the quality of children (expenditures on a child's well-being such as health or education) and consumption goods. Parents provide the same quality for all their children. The quality and quantity of children enter multiplicatively in the budget constraint of the household through the total expenditures on children. Overall expenditures on children tend to increase with income, which indicates that children are normal goods. An increase in the quantity of children raises the shadow price of the quality of children and vice versa. For example, an increase in the number of children raises the cost of providing more quality for each child because there are more children. This explains why the quantity and quality of children interact more closely than any other random pair of goods, even without assuming that both are close substitutes (Becker and Lewis 1973). If the income elasticity of demand for quality of children is higher than that of quantity, rising income increases the optimal ratio of quality to quantity of children. This implies a rise in the relative cost of an additional child relative to quality and can lead both to higher quality per child and fewer children. The income effect on fertility may be offset by the substitution effect induced by the increase in the shadow price of an additional child. If a high average education in a society generates positive externalities that boost the returns to each individual's human capital, the quality–quantity trade-offs are strengthened as families invest more heavily in each of the children (Becker et al. 1990)

Labour Market Participation and Fertility

Household production models in which both consumption and production decisions are jointly analysed provide a second major explanation for

the decrease in the number of children per woman. These models spell out how labour supply decisions are related to choices in both the timing and the level of fertility.

As a result of economic development since the Industrial Revolution, capital intensity in production increased and, further, the emphasis in activities that require physical strength diminished with the gradual shift from manufacturing towards services. This technological transformation pushed upwards the demand for activities where women have a comparative advantage and increased the relative wage of women (Galor and Weil 1996). Developed countries therefore experienced a massive entry of women into the labour market, with average female labour force participation rates climbing from 41% in 1960 to 64% by the late 1990s.

Childbearing is time-intensive relative to other activities and its associated opportunity cost can be measured by the potential wages of the mother. While increases in men's work mainly entail an income effect that increases the demand for children, increases in women's wages give rise to a combination of income and substitution effects as they result in an increase in the cost of a child relative to other goods (Mincer 1963). Accordingly, women with high potential wages may restrict their fertility and trade off children for less time-demanding alternatives if the substitution effects are important (Becker 1991). An alternative to this is the purchase of childcare services in the market. This may lessen the substitution effect and the net impact of higher wages may even turn positive (Ermisch 1989). In Scandinavian countries, for example, where publicly provided childcare is abundant, work–family tradeoffs are diminished. As women's wages have increased across the developed world, however, the cost of childcare outside the home has also risen because its provision is intensive in woman's work.

The original fertility models are static since all life-cycle choices are made at the beginning of the parent's lifetime without assuming any uncertainty. Later models emphasize the sequential nature of these decisions and incorporate stochastic shocks to the household – for example, contraceptive failure (Heckman and Willis 1976). Moffit (1984) explores how, in addition to current wage losses, lower experience and skill depreciation from career interruption may result in permanent wage gaps between women with different childbearing patterns. (His model supports the prediction that women with strong preferences for children may self-select into occupations with low wage-growth prospects: Mincer and Polachek 1974.) Hotz et al. (1997) offer a good review of dynamic models of fertility.

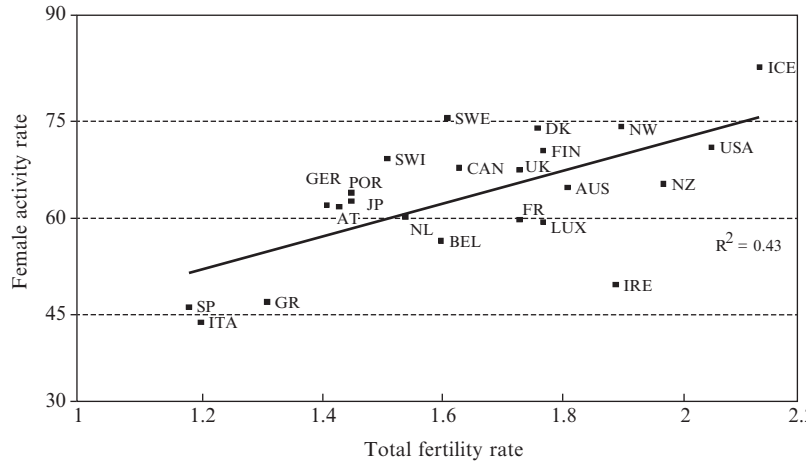
F

Family Planning

The decrease in the cost of contraception is a third important factor that facilitates the family choices discussed in the previous models. Widespread access to the birth control pill since the late 1960s had two important effects on women's careers and fertility behaviour in developed nations (Goldin and Katz 2002). First, it promoted women's career investment by reducing pregnancy risk while allowing women to remain sexually active. Second, it had an indirect impact through a social multiplier effect: the overall delay of marriage produced a thicker marriage market for career women and increased their likelihood of finding a suitable spouse later in life. Accordingly, in the United States from around 1970, more women entered professional schools and delayed marriage.

Differences Across Countries

As expected from the household production models, during the 1960s and the 1970s fertility was lower in the countries where women had entered the labour market more rapidly. Surprisingly, as female labour participation kept growing the cross-country negative relationship between fertility and labour force participation reversed by the mid-1980s. As shown in Fig. 1 for the mid-1990s, those countries with the lowest levels of female labour force participation – such as



Fertility in Developed Countries, Fig. 1 Female activity rate and total fertility rate in developed countries, 1996 (*Sources*: OECD, *Employment Outlook*, and Council of Europe)

Spain, Italy and Greece – also portrayed the lowest fertility rates. Further, even if fertility differed substantially across countries at that time (Table 1), surveys indicated that the ideal family size was above replacement level and relatively homogenous. Hence, this positive correlation was probably related to the differential support to women from government policies as well as the flexibility and performance of their labour markets.

The rapid increase in persistent unemployment since the mid-1980s was contemporaneous with the sharpest fall of fertility rates and postponement of childbearing in many countries. European unemployment went up from less than 3% before 1975 to about 10% in the 1990s. By 1990, around 50% of those unemployed in the European Union had been out of work for more than 12 months. Within the standard microeconomic model of fertility the associated fall in current opportunity costs (in terms of forgone wages) makes unemployment spells good times for childbearing (Butz and Ward 1979). Still, job loss impairs human capital accumulation and, with it, the future prospects of employment, particularly of young workers with low labour market experience. Among the employed, (temporary) withdrawals from the labour market associated with maternity have a similar effect. High and long-lasting unemployment intensifies the relevance of the latter and negative income effects can reduce fertility, as happened during the Great Depression (Becker 1991). Individuals may want to secure an adequate level of human capital (experience or education) before starting a family, and so the attractiveness of an early childbearing strategy declines. Since the 1980s, fertility postponement was more important in countries where joblessness was more prevalent and persistent – particularly among women – such as those in southern Europe (Adsera 2005).

The extent of the negative impact of unemployment, however, is related to the labour regulation and types of contractual arrangements available in each labour market. The rapid feminization of the labour force in southern Europe, where traditionally there was low female participation, collided with rigid labour market institutions that favoured traditional full-time male employment and limited the availability of part-time positions for women (Adsera 2005). In addition, the expansion of temporary contracts among young workers after partial labour reforms were passed in the late 1980s exacerbated those problems. By contrast, fertility rates are among the highest in countries with high female participation and either a flexible regulation and access to part-time employment (and low unemployment), such as the United States, the United Kingdom, or the Netherlands, or abundant public sector employment (mostly tenured jobs with features that make childbearing and participation compatible), as in the Nordic countries and France.

Changes in Values and the 'Second Demographic Transition'

Changes in values as well as secularization have long been considered independent causes of recent demographic adjustments. The fall in period fertility has been coupled with a set of changes to childbearing behaviour and family formation in most Western countries that demographers characterize as a 'second demographic transition' (Van de Kaa 1987). The most relevant features of the second demographic transition are a reduction in fertility, extensive use of modern contraceptive methods, increases in mean age at marriage and age at first birth, together with rises in extra-marital childbearing, cohabitation and divorce. In 2003, around one-third of births in developed countries occurred out of wedlock, but cross-country differences remained substantial. The share of births outside marriage ranged from just under 5% in Greece to 63% in Iceland and 56% in Sweden. At the core of these changes are an accentuation of individual autonomy, the abandonment of traditional religious beliefs, and a decline in sentiments of religiosity among individuals (Lesthaeghe and Surkyn 1988). This transformation, which was already under way in most of Western Europe during the 1970s, became increasingly widespread in southern Europe from the middle of the 1980s.

Future Implications

These demographic transformations have progressively moved to the centre of public debate both because of their social implications and the challenge they pose to the sustainability of welfare states in Western economies. As women continue to enter the labour force, labour market institutions need to adapt to minimize the trade-offs connected with childbearing to encourage fertility. In the absence of massive migration flows, smaller future cohorts facing improved economic conditions thanks to lower pressure in labour and housing markets may increase their fertility, as predicted by Easterlin's model (1975). However, since this would take place only in the long run, fertility rates are not likely to rebound to the replacement level in the near future (Bongaarts 2001). In the meantime, recent data from the Eurobarometer shows that the ideal number of children has been decreasing for women aged between 20 and 34 since the late 1980s across the European Union. The average is just above 2.1, but for the first time, some countries such as Germany and Austria already portray below-replacement desired fertility (Goldstein et al. 2003).

F

See Also

- ▶ Demographic Transition
- ▶ Easterlin Hypothesis
- ▶ Family Economics
- ▶ Human Capital, Fertility and Growth
- ▶ Labour Market Institutions
- ▶ Population Ageing

Bibliography

Adsera, A. 2005. Vanishing children: From high unemployment to low fertility in developed countries. *American Economic Review* 95: 189–193.

Becker, G.S. 1991. *A treatise on the family*, enlarged edn. Cambridge, MA: Harvard University Press.

Becker, G.S., and H.G. Lewis. 1973. On the interaction between quantity and quality of children. *Journal of Political Economy* 81(Suppl.): S279–S288.

Becker, G., K. Murphy, and R. Tamura. 1990. Human capital, fertility and economic growth. *Journal of Political Economy* 98: S12–S37.

Bongaarts, J. 2001. Fertility and reproductive preferences in post-transitional societies. *Population and Development Review* 27(Suppl.): 260–281.

Butz, W.P., and M.P. Ward. 1979. The emergence of countercyclical U.S. fertility. *American Economic Review* 69: 318–328.

Easterlin, R. 1975. An economic framework for fertility analysis. *Studies in Family Planning* 7: 54–63.

Ermisch, J. 1989. Purchased child care, optimal family size and mother's employment. *Journal of Population Economics* 2: 79–102.

Galor, O., and D. Weil. 1996. The gender gap, fertility and growth. *American Economic Review* 86: 374–387.

Goldin, C., and L. Katz. 2002. The power of the pill: Oral contraceptives and women's career and marriage decisions. *Journal of Political Economy* 110: 730–770.

Goldstein, J.R., W. Lutz, and M.R. Testa. 2003. The emergence of sub-replacement family size ideals in Europe. *Population Research and Policy Review* 22: 479–496.

Heckman, J., and R. Willis. 1976. Estimation of a stochastic model of reproduction: An econometric approach. In *Household production and consumption 40*, ed. N. Terleckyj. New York: Columbia University Press.

Hotz, V.J., J.A. Klerman, and R. Willis. 1997. The economics of fertility in developed countries. In *Handbook of population and family economics*, ed. M.R. Rosenzweig and O. Stark. Amsterdam: Elsevier.

Kohler, H.P., F. Billari, and J.A. Ortega. 2002. The emergence of lowest-low fertility in Europe during the 1990s. *Population and Development Review* 28: 599–639.

Lee, R. 2003. The demographic transition: Three centuries of fundamental change. *Journal of Economic Perspectives* 17(4): 176–190.

Lesthaeghe, R., and J. Surkyn. 1988. Cultural dynamics and economic theories of fertility change. *Population and Development Review* 14: 1–45.

Mincer, J. 1963. Market prices, opportunity costs and income effects. In *Measurement in economics: Studies in mathematical economics in Honor of Yehuda Grunfeld*. Stanford: Stanford University Press.

Mincer, J., and S. Polachek. 1974. Family investments in human capital: Earnings of women. *Journal of Political Economy* 82(Suppl.): S76–S108.

Moffit, R. 1984. Optimal life-cycle profiles of fertility and labor supply. *Research in Population Economics* 5: 29–50.

Van de Kaa, D. 1987. Europe's second demographic transition. *Population Bulletin* 42: 1–57.

Willis, R. 1973. A new approach to the economic theory of fertility behavior. *Journal of Political Economy* 81(Suppl.): S14–S64.

Fertility in Developing Countries

T. Paul Schultz

Abstract

The associations between fertility and outcomes in the family and society have been treated as causal, but this is inaccurate if fertility is a choice coordinated by families with other life-cycle decisions, including labour supply of mothers and children, child human capital, and savings. Estimating how exogenous changes in fertility that are uncorrelated with preferences or constraints affect others depends on our specifying a valid instrumental variable for fertility. Twins have served as such an instrument and confirm that the cross-effects of fertility estimated on the basis of this instrument are smaller in absolute value than their associations.

Keywords

Aging and retirement; Birth control; Capital accumulation; Child care; Child labour; Child mortality; Collective models of the household; Comparative advantage; Demographic transition; Economic growth; Family planning; Fertility in developing countries; Heckscher, E. F.; Home production; Human capital; Life-cycle savings; Malthus, T. R.; Microcredit; Mortality; Overpopulation; Poor Law; Population growth; Precautionary insurance; Time use; Value of time; Women's work and wages

JEL Classifications

J13

Fertility is a choice by parents involving a life-cycle claim on their resources, from which they may receive satisfaction as consumers and benefit as producers from children's labour and care-giving support. In addition, fertility may be the source of externalities that affect members of society other than the decision-making parents, in which case society may view fertility as a legitimate issue for social policy. To forecast fertility and the conditions under which public policies might be justified to modify fertility, economists require a basic understanding of its determinants as well as social consequences. In approaching this topic from the perspective of low-income countries today, the ideas of Malthus remain influential. He argued that population growth caused by high fertility erodes the welfare and productivity of workers, and thus social policy which fostered greater fertility, such as the English Poor Law, contributed to 'overpopulation'. Before considering how these spillover effects of fertility might be identified, an overview of historical thinking about the demographic–economic system may help to indicate the context in which Malthus's thinking was relevant to pre-industrial Europe, and how

modern economics has extended his thinking to fertility as a lifetime choice of parents related to their time allocation and accumulation of human and physical capital.

Malthus's Framework for the Pre-industrial Demographic–Economic Equilibrium

The determinants of fertility have engaged the interest of economists for some time. Adam Smith (1776) noted families were larger in settings where labour was scarce and child labour was especially valuable to parents, as in North America with its abundant land. Smith recognized that child mortality was higher among the poor, especially among those who were dependent on charity (for example, the Poor Laws). However, Malthus (1798) viewed fertility not as an individual choice but as an outcome of social institutions, because he did not think birth control was effective. He thought fertility was governed by the economic requirements society placed on a couple before allowing them to marry. Once married, the 'constant passion of the sexes' would lead in unregulated fashion to fertility. Society therefore restricted entry into marriage to those with favourable prospects for a livelihood or the income and assets to support the children that were expected to follow from the union. Over his lifetime, Malthus accumulated corroborating evidence on fertility, population growth and economic growth. Historians have since added to Malthus's evidence, confirming that Europe exhibited a late median age at marriage for a woman in her mid-twenties. This delay in childbearing led European women to have four or five births over their lifetime, rather than the six or seven if they had married five years earlier. Given the short life expectancy in pre-industrial Europe of about 35–40 years, this restrained level of fertility diminished substantially the resulting rate of population growth, except at frontiers of settlement where labour was scarce, land abundant, and marriage consequently early.

Heckscher (1963) thought Malthus's framework was relevant to Sweden. With the Swedish church's good records of marriages, births and deaths, and the Swedish king's need to estimate crop yields (for the purposes of taxation), annual time series for Sweden after 1720 appear accurate and show a positive covariation in marriage and fertility with good crop years, and shortfalls in marriage and subsequently fertility following poor crop years. Temperature and rainfall data available for Sweden after 1750 allow later analysts to incorporate this exogenous variation in weather and employ vector autoregression to estimate weather-driven Malthusian cycles in wages, fertility, as well as mortality (Eckstein et al. 1985).

Working with French and Swiss parish registries of marriage, births, and deaths, Louis Henry (1972), the demographer, found evidence that couples exhibited a 'natural' rate of childbearing after marriage, until they eventually began to increase the intervals between their births after later parities, if economic conditions became less favourable. The emergence of this form of parity-specific application of birth control over the life-cycle of marriages was interpreted by Coale (1973) as an indicator of the onset of the 'demographic transition', when cultural restraints on fertility evolved from 'natural' proximate determinants to controlled 'modern' reproductive behaviour relying primarily on birth control.

Parish registries were then sampled from England from 1541 to 1871 by Wrigley and Schofield (1981) to further investigate the Malthusian framework. Lee (1981) found that increases in marriage and birth rates were related to good weather and resulting declines in the price of wheat, as Malthus would have expected. But only about half of the covariation in weather/prices and annual birth rates is due to the fluctuations in first births that follow in the wake of variations in marriage. The other half is explained by variation in the length of inter-birth intervals. The latter finding casts doubt on Malthus's view that in this pre-industrial period couples did not exercise fertility choices within marriage. This spacing of births in response to economic wage cycles implied that the adoption of parity-specific birth control may not have been a cultural

innovation, as assumed by Coale, but a customary form of individual behaviour adopted when additional births were unwanted. Some couples in pre-industrial societies appear able and willing to practice effective birth control when motivated economically. Fertility is thus to some degree a voluntary choice variable within marriage even in pre-industrial societies.

As the Industrial Revolution progressed in Europe and real wages increased, fertility nonetheless began to decline widely by the end of the 19th century. The Malthusian framework needed to be amended further to fit this experience in Europe and be applicable to low-income countries after 1960 as new methods of family planning were disseminated in the world and fertility fell despite modern economic growth. How was the secular decline in fertility to be explained in the face of rising personal incomes? The decline in child mortality, which gathered speed after 1870, reduced the need for parents to have extra births to replace the one out of five who might have at earlier times died from childhood diseases and infections. Parents might also scale back their demand for 'insurance' births motivated to reduce the likelihood that a couple would sustain above average child losses (Schultz 1981). Becker (1960) proposed that the relative price of rearing children increased over time, causing the decline in parents' demand for children. Mincer (1963) hypothesized that an increase in women's wages increased a couple's opportunity cost of having children, raising the shadow price of children. He argued that the rise in female labour-force participation and the decline in fertility were both caused by conditions increasing women's wages relative to other consumer prices and men's wages. These empirical patterns in the United States were soon replicated in other high-income countries.

Changing the relative prices of outputs of the economy is one possible source of variation in women's wages relative to men's that could explain changes in fertility. Men's labour in European agriculture was critical for plowing and producing food grains, whereas women specialized in home production as domestic servants and wives and to some degree in animal husbandry and the production of dairy commodities. Consequently, changing scarcity of grains relative to livestock and dairy product contributed to swings in the relative wages of men and women in Europe. The secular decline in international grain prices relative to dairy and livestock prices in the latter half of the 19th century was unprecedented due to the opening of new lands at the frontiers of European settlement in the United States and Russia, and contributed along with changes in production technologies to the rise in women's agricultural wages relative to men's in northern Europe and to the decline in fertility. Swedish historical data by region document after 1860 the fall in world grain prices, the associated increase in the wages of women relative to men, and the secular fall in fertility, when other developments are controlled for (Schultz 1985).

Another factor credited with reducing fertility is the improvement in birth control technology, which reduced the monetary and psychic cost of limiting births, and provided techniques controlled by women, which were independent of sex. The major advances in technology occurred in the 1960s with the introduction of oral steroids (the pill) and the intra-uterine device (IUD), followed by further refinements in their delivery systems. Traditional mechanisms for population control such as abortion, infanticide, coitus interruptus, and condoms have nonetheless allowed individuals to adjust their family size and affect population growth in various periods and parts of the world, well before the advent of these modern means of birth control. Although they may have facilitated the later demographic transition, these birth control technologies do not appear to have been necessary.

Microeconomic Models of Fertility Behaviour

Willis (1973) adapted a comparative advantage trade model to the household lifetime fertility choice problem, wherein women's education was assumed to enhance women's productivity only in the market, and thereby increase the relative price of home production and decrease their

demand for fertility. In his economic treatise on the family, Becker (1981) assigns a central role to market/non-market specialization of spouses in the household, with childbearing and rearing being the dominant non-market production activity traditionally performed by women.

To place more structure on fertility choices, Becker (1960, 1981) and Willis (1973) hypothesize that parents viewed the human capital of their children (child quality) as a substitute for their number of children (child quantity). If this were the case, then by definition income-compensated cross-price effects should be positive between child quantity and quality. In other words, increasing the price of children, for example by reducing the cost of birth control, would directly decrease fertility and indirectly increase the demand for child quality (with income held constant). Conversely, increasing the wage returns to schooling in the labour market would directly increase the demand for schooling and indirectly decrease the demand for births. Becker and Lewis (1974) postulate further that the income elasticity of demand for child quality exceeded the positive income elasticity for child quantity, which could account for the paradoxical decline in fertility with growth in income, without having to assume that children (quantity) are an 'inferior' good for which income effects are negative, or to show increases in women's value of their time in the modern economy caused the decline in their fertility.

The decline in fertility by half in high-income countries during the 20th century brought population growth to a halt in many of these countries. The decline in fertility by more than half in low-income countries in 40 years (1965–2005) is not yet comprehensively accounted for, although demographers are agreed that these trends in fertility are irreversible and the size of the world's population will stabilize later in the 21st century. How much does each of these conceptually distinct factors economists have described explain of this remarkable decline in fertility? I do not yet find a consensus on how to weight these factors in explaining cohort fertility. What fraction is due to an exogenous decline in mortality, the decline in the relative value of child labour, the increase in the value of women's time used in child care and the related increase in their empowerment, the increase in returns to schooling children, the greater income elasticities of demand for child quality than for quantity, and finally the improvements in birth control technology?

Identifying the Effect of Fertility on the Welfare of Families and Society

The policy-relevant externalities of fertility could arise at the aggregate level or in terms of substitution effects within families. Malthus assumed that fertility added to subsequent generations of workers, which reduced their wages and also changed the age composition of the population. But empirical evidence for these aggregate effects of fertility has not led to a consensus on their importance for today's low-income countries (National Research Council 1986). At the microeconomic level of the family, fertility is found to be closely associated with other life-cycle choices by parents, including the share of time women allocate to the market economy, the investments parents make in the human capital of each of their children, and perhaps the savings out of income they accumulate in physical capital, possibly for old age support or precautionary insurance. But to assess the magnitude of these cross-effects of fertility, researchers must first specify an exogenous factor (not a choice variable within the orbit of the family) that affects fertility but leaves other constraints on the family life-cycle choices and outcomes unaffected and is unrelated to parent preferences (Schultz 2007). In other words, an exclusion restriction or a valid instrumental variable is needed to account for some part of the variation in fertility that is independent of parent preferences and family life-cycle economic constraints. Otherwise, these cross-effects observed at the family level may not be causal and cannot be expected to occur when population policies reduce (or increase) fertility.

Twins are proposed by Rosenzweig and Wolpin (1980, 2000) as a 'shock' to the quantity of children that is uncorrelated with parent preferences or unobserved determinants of other

family and child outcomes. Adjustment of investment in the schooling of other children in the family due to the occurrence of twins can then test the quantity–quality substitution hypothesis. They found support for the trade-off of quantity–quality on non-twin siblings in rural Indian households observed in 1970. A larger sample of twins collected in China provides the basis for estimating the impact of a twin on the quality of earlier- or later-born siblings, providing bounds to the magnitude of the cross effects, adjusted for substitution effects between siblings (Rosenzweig and Zhang 2006). However, when twins are an instrument for fertility, the estimated quantity–quality trade-off tends to be smaller in absolute value than when estimated by direct association, that is, ordinary least squares (OLS). This could be due to the twin instrument being weak either because it occurs for only a small fraction of births (for example, one per cent) or because the underlying causal relationship is in fact weak and appears important only in biased single-equation associations (that is, OLS). The heterogeneity in parent preferences or other unobserved determinants of behavior could inversely affect child quantity and quality (Schultz 2007).

Other studies have exploited twins as an instrument for fertility to assess how exogenous fertility affects the mother's market labour supply. These studies in highand low-income countries generally confirm that the twin instrumental variable estimate of the effect of a birth on the mother's market labour supply tends to be absolutely smaller (negative) than the OLS estimate. The Durbin–Wu–Hausman specification test rejects the exogeneity of fertility in the determination of the mother's allocation of time to market work (Schultz 2007), implying that the consistent instrumental variable estimate is preferred over the OLS estimate.

This twin-based cross effect of fertility on mothers' labour supply may help to explain how policies which reduce fertility can facilitate modern economic growth, by adding to the per capita supply of labour and increasing the human capital of future generations. Finally, if parents when they have fewer children increase life-cycle savings for their support in old age, policies that facilitate a decline in fertility could raise savings and further augment growth rates. But estimates of these three potential cross effects of fertility-reducing population policies remain currently speculative.

The other instrument commonly used to identify the consequences of fertility on the welfare of families relies on the sex composition of births, and has serious drawbacks. This variable may significantly affect parents' decisions on whether to have further children, and it may be assumed to be approximately independent of parent preferences or family constraints if there is no sex-selective abortion or infanticide. But this variable may not satisfy the criteria for a valid instrument, because the social and economic consequences of a child's sex involve many culturally distinct costs and benefits for his or her parents, such as the provision of dowries for daughters in some parts of the world. Thus, the sex composition of early births is likely to involve lifetime wealth effects for parents, in addition to affecting fertility, giving rise to many changes in family time allocation, expenditure patterns, and life-cycle savings (Rose 2000). Therefore, the sex composition of children is not an instrumental variable for estimating how parents respond to a change in their fertility due to a population policy, if income and other family constraints are held constant. Finally, it should be noted that population policies may on the one hand subsidize learning and use of birth control, or at the other extreme fix a birth quota, as in China. There is no reason to expect expanding voluntary choices in the first case will have the same effect as rationing choices in the other policy regime.

Conclusions and Research Challenges

Parents may altruistically internalize in their fertility decisions the effects of their fertility on their welfare and that of their children, including investments in child quality and lifetime savings in financial assets (Becker 1981). These parents are typically assumed to have secure property rights to their savings and access to financial institutions that minimize credit constraints.

Population policies that reduce the cost of avoiding unwanted births may also be expected to affect gender empowerment, which does not enter decisively in the unitary model of the family proposed by Becker, but emerges in various recent bargaining and collective models of the family. Women may differentially gain from improved control of reproduction, because they physically bear the health costs of having births and invest disproportionately their time in child rearing. To derive predictions on how family bargaining affects fertility or vice versa requires more context-specific assumptions. Do mothers or fathers value children more highly? Does improved birth control technology empower women to bargain for a larger share of the gains from marriage? These remain open questions for more study. Women may value children as much as men do, and use their own increases in wealth to have more. Increased unearned income owned by the wife is associated, if the husband's income is held constant, with higher fertility in Thailand but not in Brazil (Schultz 1990). Microcredit targeted to groups of women in Bangladesh increases women's earnings and increases their later fertility (Pitt et al. 1999).

In an experimentally designed family planning and health programme started in 1977 for women in rural villages of Matlab, Bangladesh, the women in villages benefiting from the programme had one fewer child by 1996 than did comparable women in comparison villages (Joshi and Schultz 2006). The programme is also associated with increased woman's health, as measured by their body mass index (weight divided by height squared), reduced child mortality before age five, and increased years of schooling of boys aged 9–14 and 15–29. More studies of these long-run consequences of population policies on fertility and other family outcomes will be needed to assess the within-family consequences of fertility and population policies. Recognition that fertility is endogenous to other family life-cycle choices challenges economists to measure these potentially important life-cycle causal connections, and thereby provide a sounder basis for evaluating how population policies affects the social allocation of resources.

See Also

- ▶ Child Labour
- ▶ Fertility in Developed Countries
- ▶ Human Capital, Fertility and Growth

Bibliography

Becker, G.S. 1960. An economic analysis of fertility. In *Demographic and economic change in developed countries*. Princeton: Princeton University Press and NBER.

Becker, G.S. 1981. *A treatise on the family*. Cambridge: Harvard University Press.

Becker, G.S., and H.G. Lewis. 1974. Interactions between quantity and quality of children. In *Economics of the family*, ed. T.W. Schultz. Chicago: University of Chicago Press.

Coale, A.C. 1973. The demographic transition reconsidered. In *Proceedings of the international population conference*, vol. 1. Liège: International Union for the Scientific Study of Population.

Eckstein, Z., T.P. Schultz, and K. Wolpin. 1985. Short run fluctuations in fertility and mortality in pre-industrial Sweden. *European Economic Review* 26: 295–317.

Heckscher, E.F. 1963. An economic history of Sweden, trans. G. Ohlin. Cambridge: Harvard University Press.

Henry, L. 1972. *On the measurement of human fertility*. Amsterdam: Elsevier.

Joshi, S., and T.P. Schultz. 2006. Family planning as a long term investment in development. Discussion paper, Economic Growth Center, Yale University.

Lee, R. 1981. Short term variation: Vital rates, prices, and weather. In *The population history of England, 1541–1871*, ed. J.A. Wrigley and R.S. Schofield. Cambridge: Harvard University Press.

Malthus, T.R. 1798. In *Essay on the principle of population*, ed. A. Flew, 1970. Harmondsworth: Penguin.

Mincer, J. 1963. Market prices, opportunity costs and income effects. In *Measurement in economics*, ed. C. Christ et al. Stanford: Stanford University Press.

National Research Council. 1986. *Population growth and economic development: Policy questions*. Washington, DC: National Academy Press.

Pitt, M., S.R. Khandkar, S.-M. McKernan, and M.A. Latif. 1999. Credit programs for the poor and reproductive behavior in low income countries. *Demography* 35: 1–21.

Rose, E. 2000. Gender bias, credit constraint and time allocation in rural India. *Economic Journal* 110: 738–758.

Rosenzweig, M.R., and K.I. Wolpin. 1980. Testing the quantity–quality fertility model: the use of twins as a natural experiment. *Econometrica* 48: 227–240.

Rosenzweig, M.R., and K.I. Wolpin. 2000. Natural 'natural' experiments in economics. *Journal of Economic Literature* 38: 827–874.
Rosenzweig, M.R., and J. Zhang. 2006. Do population control policies induce more human capital investment? Twins, birthweight, and China's 'one child' policy. Discussion Paper No. 933, Economic Growth Center, Yale University.
Schultz, T.P. 1981. *Economics of population*. Reading: Addison Wesley.
Schultz, T.P. 1985. Changing world prices, women's wages, and the fertility transition. *Journal of Political Economy* 93: 1126–1154.
Schultz, T.P. 1990. Testing the neoclassical model of family labor supply and fertility. *Journal of Human Resources* 25: 559–634.
Schultz, T.P. 2007. Population policies, fertility, women's human capital and child quality. In *Handbook of development economics*, ed. T.P. Schultz, vol. 4. Amsterdam: North-Holland.
Smith, A. 1776. In *The wealth of nations*, ed. E. Cannan, 1961. London: Methuen.
Willis, R.J. 1973. A new approach to the economic theory of fertility behavior. *Journal of Political Economy* 81(2, Part II), S14–S64.
Wrigley, J.A., and R.S. Schofield. 1981. *The population history of England, 1541–1871*. Cambridge: Harvard University Press.

Fetter, Frank Albert (1863–1949)

Murray N. Rothbard

Keywords

Austrian economics; Basing-point pricing; Böhm-Bawerck, E. von; Business cycles; Capitalization; Distribution theories; Fetter, F. A.; Productivity; Rent; Time preference; Uniform pricing; Value theories

JEL Classifications

J13

Fetter was born on 3 March 1863 in the town of Peru, Indiana, and died on 21 March 1949 in Princeton, New Jersey. He was educated at Indiana and Cornell Universities and received his doctorate in economics at the University of Halle in Germany in 1894; he spent most of his life teaching at Cornell (1901–1911) and Princeton universities (1911–1934).

In journal articles on capital, interest and rent written largely between 1900 and 1914 (Fetter 1977), and particularly in two treatises on economic principles (Fetter 1904, 1915), Fetter built upon Böhm-Bawerk and the Austrian School to develop a lucid and remarkable integrated structure of economic theory. He was able to accomplish this feat by purging economics of all traces of Ricardian or other British objectivist theories of value and distribution, in particular any differential theories of rent or productivity theories of interest.

Much of Fetter's achievement rested on his insight into the ordinary language meaning of 'rent' as simply the price of any durable good per unit time. He was then able to show that the prices of consumer goods are determined by their marginal utilities, and that these values are imputed back to determining the rental prices of factors of production by their marginal value productivity in serving consumers. The capital value, or price of the whole good (whether land, capital goods, or, Fetter might have added, the labourer under slavery) is then determined by the sum of its expected future returns, or rents, discounted by the social rate of time preference, or rate of interest. Thus, Fetter went beyond Böhm-Bawerk by arriving at a pure time preference theory of interest. Productivity and time preference are both highly important, but they have very different functions: the former in determining rents, and the latter determining the rate of interest. Thus, future rents are discounted by the rate of time preference and summed up, or 'capitalized', into their present capital value. Indeed, Fetter often called his contribution the 'capitalization theory of interest'.

Fetter presented the fullest portrayal yet attained of the time market, the market for present as against future goods, as it permeates the economic system. The time market is not only the loan market, but also exists when entrepreneurs purchase or hire discounted factors of

production (future goods) in return for money (a present good) and then reap a time or interest return when the product is later sold as a present good. Entrepreneurs earn profits, or suffer losses, as they lead the economy in the direction of a general equilibrium determined by marginal utility, marginal value productivity, and time preference.

While Fetter was led by his capitalization theory to arrive independently at the Mises–Hayek theory of the business cycle in 1927 (Fetter 1977, pp. 260–316), he virtually abandoned value and distribution theory in the last two decades of his life to concentrate on the alleged monopolistic evils of basing-point pricing. He assumed that competition requires uniform pricing of products at the mill, while uniform pricing at centres of consumption is somehow monopolistic and deserves to be outlawed (Fetter 1931). Fetter's shift of concern, coupled with a general loss of interest in economic theory in the United States between the two world wars and the continuing dominance of neo-Ricardian Marshallian theory in Britain, gravely hindered the incorporation of Fetter's notable contributions into modern economics.

Selected Works

1904. The principles of economics. New York: Century.
1915. *Economic principles.* New York: Century.
1931. *The masquerade of monopoly.* New York: Harcourt, Brace.
1977. Capital, interest and rent: *Essays in the theory of distribution, ed. M. Rothbard. Kansas City: Sheed, Andrews and McMeel.*

Bibliography

Coughlan, J.A. 1965. *The contributions of Frank Albert Fetter (1863–1949) to the development of economic theory.* Doctoral dissertation, Catholic University of America, Washington, DC.
Hoxie, R.F. 1905. Fetter's theory of value. *Quarterly Journal of Economics* 19: 210–230.

Fetter, Frank Whitson (Born 1899)

Barry Gordon

Fetter was born in San Francisco, California, in 1899. His published research is wide-ranging, including studies of inflation and international economic issues, but his most celebrated contributions are in the history of economic thought. These contributions were accorded special recognition in 1982, when he became a Distinguished Fellow of the History of Economics Society.

After gaining a first degree at Swarthmore (BA, 1920), Fetter went to Harvard (MA, 1924) and Princeton (MA, 1922; PhD, 1926). Thereafter, he taught economics at Princeton (to 1934) and at Haverford College (1934–48). In 1948 he was appointed Professor of Economics at Northwestern University and remained in that post until his retirement in 1967.

Fetter chose classical economics as the major focus of his research, in particular British economic thought from Adam Smith to John Stuart Mill. That thought he has characterized as, 'a time bomb under the citadels of the established order' (Fetter 1981, p. 31). In his view the core of classical economics was not the doctrine of laissez-faire. Rather, it was rationality, and this led economists to be concerned with questions such as religious discrimination and aristocratic privilege, as well with the freer operation of the forces of the market. They were advocates of social change on a broad front.

Given this understanding of classical economics, Fetter's work has not been confined to textual analysis of the treatises of the great theorists. In addition he has closely observed economists at work in the public forums of 19th-century Britain, as shown in his masterly overview of their interventions in Parliament (Fetter 1980). This book examines the economist's role in debates concerning not only trade, working conditions, business practice, taxation and other economic matters, but also on such

issues as education, church-state relations, civil rights and parliamentary reform.

Another forum for economists in the first half of the 19th century was that provided by influential periodicals such as the *Edinburgh Review*, the *Westminster Review* and *Blackwood's*. This facet of contemporary economic debate he explored in a series of pioneering papers (Fetter 1953, 1958, 1960, 1962, 1965). Special mention is also due his work on the development of thought relating to monetary and banking policy in Great Britain (Fetter 1955, 1965, 1973), which brings the modern reader into intimate contact with the institutions, personalities and conceptual divisions that were crucial in the evolution of a powerful monetary orthodoxy.

The contributions of Fetter are informed by the conviction that a grasp of history is a vital element in the intellectual equipment of those who would make economic judgements.

Selected Works

1931. *Monetary inflation in Chile*. Princeton: Princeton University Press. Spanish trans. as *La inflación monetariau en Chile*. Santiago: University of Chile, 1937.

1942. The life and writings of John Wheatley. *Journal of Political Economy* 50: 357–376.

1953. The authorship of economic articles in the *Edinburgh Review*, 1802–47. *Journal of Political Economy* 61: 232–259.

1955a. *The Irish pound, 1797–1826*. London: Allen and Unwin.

1955b. Does America breed depressions? *Three Banks Review* 27: 28–41.

1957. (ed.) *The economic writings of Francis Horner*. London: London School of Economics and Political Science.

1958. The economic articles in the *Quarterly Review* and their authors, 1809–1852. *Journal of Political Economy* 66 Pt I: 47–64; Pt II: 154–170.

1959. The politics of the Bullion Report. *Economica* 26: 99–120.

1960. The economic articles in *Blackwood's Edinburgh Magazine*, and their authors, 1817–1853. *Scottish Journal of Political Economy* 7 Pt I: 85–107; Pt II: 213–231.

1962a. Robert Torrens: Colonel of Marines and political economist. *Economica* 29: 152–165.

1962b. Economic articles in the *Westminster Review* and their authors, 1824–51. *Journal of Political Economy* 70: 570–596.

1964. (ed.) *Selected economic writings of Thomas Attwood*. London: London School of Economics and Political Science.

1965a. *Development of British Monetary orthodoxy, 1797–1875*. Cambridge, MA: Harvard University Press.

1965b. Economic controversy in the *British Review*, 1802–1850. *Economica* 32(128): 424–437.

1968. The transfer problem: Formal elegance or historical realism? In *Essays in money and banking in honour of R.S. Sayers*, ed. C.R. Whittlesey and J.S.G. Wilson, 63–84. Oxford: Oxford University Press.

1969. The rise and decline of Ricardian economics. *History of Political Economy* 1(1): 67–84.

1973. (With D. Gregory.) *Monetary and financial policy*. Dublin: Irish University Press.

1975. The influence of economists in Parliament on British legislation from Ricardo to John Stuart Mill. *Journal of Political Economy* 83(5): 1051–1064.

1980. *The Economist in parliament, 1780–1868*. Durham: Duke University Press.

1981. Are economists of any use? *History of Economics Society Bulletin* 3.

Feudalism

Robert Brenner

Keywords

Capitalist property relations; Class; Colonization; Demesne production; Diversified production; Dobb, M. H.; Exchange; Feudalism; Marx, K. H.; Peasant economy; Peasants;

Population growth; Pre-capitalist property relations; Subsistence; Sweezy, P. M.; Vassalage

JEL Classifications
B31

Modern discussions of feudalism have been bedevilled by disagreement over the definition of that term. There are three main competing conceptualizations. (1) Feudalism refers strictly to those social institutions which create and regulate a quite specific form of legal relationship between men. It constitutes a relationship in which a freeman (vassal) assumes an obligation to obey and to provide, primarily military, services to an overlord who, in turn, assumes a reciprocal obligation to provide protection and maintenance, typically in the form of a fief, a landed estate to be held by the vassal on condition of fulfilment of obligations (Bloch 1939–40). (2) Feudalism refers, more broadly, to a form of government or political domination. It is a form of rule in which political power is profoundly fragmented geographically; in which, even within the smallest political units, no single ruler has a monopoly of political authority; and in which political power is privately held, and can thus be inherited, divided among heirs, given as a marriage portion, mortgaged, and bought and sold. Finally, the armed forces involve, as a key element, a heavy armed cavalry which is secured through private contracts, whereby military service is exchanged for benefits of some kind (Strayer 1965; Ganshof 1947). (3) Feudalism refers to a type of socio-economic organization of society as a whole, a mode of production and of the reproduction of social classes. It is defined in terms of the social relationships by which its two fundamental social classes constitute and maintain themselves. Specifically, the peasants, who constitute the overwhelming majority of the producing population, maintain themselves by virtue of their possession of their full means of subsistence, land and tools, so require no productive contribution by the lords to survive. This possession is secured by means of the peasants' collective political organization into self-governing communities, which stand as the ultimate guardian of the individual peasants' land. As a result of the peasants' possession and their consequent economic independence, mere ownership of property cannot be assumed to yield an economic rent; in consequence, the lords are obliged to maintain themselves by appropriating a feudal levy by the exercise of *extra-economic* coercion. The lords are able to extract a rent by extra-economic coercion only in consequence of their political self-organization into lordly groups or communities, by means of which they exert a degree of domination over the peasants, varying in degree from enserfment to mere tribute taking (Marx 1894; Dobb 1946).

Though often thought to be in conflict, these conceptions are not only complementary but in fact integrally related to one another. While the lords' very existence as lords was based, as Marxists correctly insist, upon their appropriating a rent from the peasantry by extra-economic coercion, their capacity actually to exert such force in the rent relationship depended upon from their ability to construct and maintain the classically political ties of interdependence which joined overlord to knightly follower and thereby constituted the feudal groups which were the ultimate source of the lords' power. Conversely, while feudal bonds of interdependence were constructed, as the Weberians emphasize, to build highly localized governments capable at once of waging warfare, dispensing justice and keeping the peace, the *raison d'être* of the mini-states thus created was to constitute the dominant class of feudal society by establishing the instruments for extracting, redistributing and consuming the wealth upon which this class depended for their maintenance and reproduction. State and ruling class were thus two sides of the same coin. The distinctive ties which bound man to man in feudal society (not only the relations of vassalage strictly speaking, but also the more loosely defined associations structured by patronage, clientage, and family) constituted the building blocks, at one and the same time, for the peculiarly fragmented, locally based and politically competitive character of the feudal ruling class and for the peculiarly

particularized nature of the feudal state. It was the lords' feudal levies which provided the material base for the feudal polity. It was the parcellized character of the feudal state, itself the obverse side of the decentralized structure of lordship through which rent was appropriated from the peasantry, which thus created the basic opportunities, set the ultimate limits and posed the fundamental problems for the lords' reproduction as a ruling class.

The Origins of Feudalism

The rise of feudalism was conditioned by an extended process of political fragmentation within the old Carolingian Empire. This is understandable, in part, in terms of a tendency to decentralization inherent in patrimonial rule. The patrimonial lord, to maintain his following, had, paradoxically, to provide his followers with the means to establish their independence from him. He could counteract their tendency to assert their autonomy through successful warfare and conquest, in which the followers found it worth their while to continue to submit to his authority. But in the absence of such profitable aggression, the followers had every incentive to assert their independence. It was in this way that the devolution and dissolution of more centralized forms of authority took place within the Carolingian Empire during the 9th and 10th centuries, as the Franks and their followers ceased to be conquerors, following a long period in which the empire had expanded. Fragmentation was hastened by the contemporaneous invasions of the Northmen, Saracens and Magyars. Effective authority fell, successively, from the king to his princes, to the counts and, ultimately, to local castleholders and even manorial lords, as the newly emerging, highly localized rulers turned their pillaging from foreign enemies to the local population (Weber 1956; Duby 1978, pp. 147 ff).

Feudalism originally took shape in the early part of the 11th century in many parts of western Europe, including much of France, northern Italy and western Germany. Feudal rule was first constituted through the formation of lordly political groups, initially organized around a castle and led by the castellan. The castellan's power was derived from his knightly followers. The knights possessed military training, fought on horseback wearing (increasingly elaborate) coats of armour, often lived in the castle, and, from around the second third of the 11th century, tended to be bound to the castellan through ties of vassalage. The castellan's hegemony was manifested in his capacity to exert the right of the ban over his district – whose outer limits were usually no more than half a day's ride from the central fortress. The right of the ban, traditionally in the hands of the early medieval kings and the direct expression of their authority, allowed the castellan, above all, to extract dues from the peasant households within his jurisdiction, as well as to dispense justice and keep the peace. Although the surrounding lesser lords were usually tied to a castellan, in some cases they retained their full independence, not only collecting feudal rents derived from their authority over their tenants, but imposing taxes and exerting justice within their manorial mini-jurisdictions. In any case, all these lords confirmed their membership in the dominant class by claiming exemption from fiscal exactions: freedom under feudalism thus took the form of privilege. The peasants' unfreedom in some cases originated from their ancestors' having formally commended themselves to their lord; that is, their having subjected themselves to his domination in exchange for his assuring their safety. But, with the crystallization of feudal domination, it simply expressed the lords' having appropriated the right to extort protection money from them. The peasants' unfreedom was thus defined and constituted precisely by their subjection to arbitrary levies (Duby 1973, 1978).

The feudal economy was thus structured, on the one hand, by a form of pre-capitalist property relations in which the individual peasant families, as members of a village community, *individually possessed* their means of reproduction. This contrasted with other pre-capitalist property forms in which the village community itself was the possessor (or more of one). On the other hand, under feudalism, the individual lords reproduced themselves by *individually appropriating* part of the peasants' product, backed up by localized

communities of lords connected by various sorts of political bond, classically vassalage. This contrasted with other pre-capitalist property systems, in which the community, or communities, of lords appropriated the peasants' product collectively (as a tax) and shared out the proceeds among the community's, or communities', members.

Feudal Property Relations and the Forms of Individual Economic Rationality

The fundamental feudal property relationships of peasant possession and of lordly surplus extraction by extra-economic compulsion shaped the long-term evolution of the feudal economy. This was because these relationships were systematically maintained by the conscious actions of communities of peasants and of lords and thus constituted relatively inalterable constraints under which individual peasants and lords were obliged to choose the patterns of economic activity most sensible for them to adopt in order to maintain and improve their condition. The potential for economic development under feudalism was thus sharply restricted because both lords and peasants found it in their rational self-interest to pursue individual economic strategies which were largely incompatible with, if not positively antithetical to, specialization, productive investment and innovation in agriculture.

First, and perhaps most fundamental, because both lords and peasants were in full possession of what they needed to maintain themselves as lords and peasants, they were free from the *necessity* to buy on the market what they needed to reproduce, thus freed from dependence on the market and the necessity to produce for exchange, and thus exempt from the requirement to sell their output competitively on the market. In consequence, both lords and peasants were free from the necessity to produce at the socially necessary rate so as to maximize their rate of return and, in consequence, relieved of the requirement to cut costs so as to maintain themselves, and so of the necessity constantly to improve production through specialization and/or accumulation and/or innovation. Feudal property relations, in themselves, thus failed to *impose* on the direct producers that relentless drive to improve efficiency so as to survive, which is the *differentia specifica* of modern economic growth and required of the economic actors under capitalist property relations in consequence of their subjection to production for exchange and economic competition.

Absent the necessity to produce so as to maximize exchange values and in view of the underdeveloped state of the economy as a whole, the peasants tended to find it most sensible actually to deploy their resources so as to ensure their maintenance by producing directly the full range of their necessities; that is, *to produce for subsistence.* Given the low level of agricultural productivity which perforce prevailed, harvests and therefore food supplies were highly uncertain. Since food constituted so large a part of total consumption, the uncertainty of the food market brought with it highly uncertain markets for other commercial crops. It was therefore rational for peasants to avoid the risks attached to dependence upon the market, and to do so they had to diversify rather than specialize, marketing only physical surpluses. In fact, beyond their concern to minimize the risk of losing their livelihood, the peasants appear to have found it desirable to carry out diversified production simply because they wished to maintain their established mode of life – and, specifically, to avoid the subjection to the market which production for exchange entails, and the total transformation of their existence which that would have meant.

To make possible ongoing production for subsistence, the peasants naturally aimed to maintain their plots as the basis for their existence. To ensure the continuance of their families into the future, they also sought to ensure their children's inheritance of their holdings. Meanwhile, they tended to find it rational to have as many children as possible, so as to ensure themselves adequate support in their old age. The upshot was relatively large families and the subdivision of plots on inheritance.

Like the peasants, the lords occupied a 'patriarchal' position, possessing all that they needed

to survive and thus freed of any necessity to increase their productive capacities. Moreover, even to the extent they wished, for whatever reason, to increase the output of their estates, the lords faced nearly insuperable difficulties in accomplishing this by means of increasing the productive powers of their labour and their land. Thus, if the lords wished to organize production themselves, they had no choice but to depend for labour on their peasants, who possessed their means of subsistence. But precisely because the peasants were possessors, the lords could get them to work only by directly coercing them (by taking their feudal rent in the form of labour) and could *not* credibly threaten to 'fire' them. The lords were thereby deprived of perhaps the most effective means yet discovered to impose labour discipline in class-divided societies. Because the peasant labourers had no *economic* incentive to work diligently or efficiently for the lords, the lords found it extremely difficult to get them to use advanced means of production in an effective manner. They could force them to do so only by making costly unproductive investments in supervision.

In view of both the lords' and the peasants' restricted ability effectively to allocate investment funds to improved means of production to increase agricultural efficiency, both lords and peasants found that the only really effective way to raise their income via productive investment was by opening up new lands. Colonization, which resulted in the multiplication of units of production on already existing lines, was thus the preferred form of productive investment for both lords and peasants under feudalism.

Beyond colonization and the purchase of land, feudal economic actors, above all feudal lords, found that the best way to improve their income was by forcefully *redistributing* wealth away from the peasants or from other lords. This meant that they had to deploy their resources (surpluses) towards building up their *means of coercion* by means of investment in military men and equipment, in particular to improve their ability to fight wars. A drive to *political accumulation,* or state building, was the feudal analogue to the capitalist drive to accumulate capital.

The Long-Term Patterns of Feudal Economic Development

Feudal property relations, once established, thus obliged lords and peasants to adopt quite specific patterns of individual economic behaviour. Peasants sought to produce for subsistence, to hold on to their plots, to produce large families and to provide for their families' future generations by bequeathing their plots. Both lords and peasants sought to use available surpluses funds to open new lands. Lords directed their resources to the amassing of greater and better means of coercion. Generalized on a society-wide basis, these patterns of individual economic action determined the following developmental patterns, or laws of motion, for the feudal economy as a whole:

(i) *Declining productivity in agriculture* (Bois 1976; Hilton 1966; Postan 1966)

The generalized tendency to adopt production for subsistence on the part of the peasantry naturally constituted a powerful obstacle to commercial specialization in agriculture and to the emergence of those competitive pressures which drive a modern economy forward. In so doing, it also posed a major barrier to agricultural improvement by the peasantry, since a significant degree of specialization was required to adopt almost all those technical improvements which would come to constitute 'the new husbandry' or the agricultural revolution (fodder crops, up-and-down farming, and so on). In addition, production aimed at subsistence and the maintenance of the plot as the basis for the family's existence posed a major barrier to those rural accumulators, richer peasants and lords who wished to amass land or to hire wage labour, since the peasants would not readily part with their plots, which were the immediate bases for their existence, unless compelled to do so; nor could they be expected to work for a wage unless they actually needed to.

Further counteracting any drive to the accumulation of land and labour was the tendency on the part of the possessing peasants to produce large families and subdivide their holdings among their

children. The peasants' parcellization of plots under population growth tended to overwhelm any tendency towards the build-up of large holdings in the agricultural economy as a whole, further reducing the potential for agriculture improvement.

Finally, individual peasant plots were, most often, integrated within a village agriculture which was, in critical ways, controlled by the community of cultivators. The peasant village regulated the use of the pasture and waste on which animals were raised, and the rotation of crops in the common fields. Individual peasants thus tended to face significant limitations on their ability to decide how to farm their plots and thus, very often, on their capacity to specialize, build up larger consolidated holdings, and so forth.

To the extent that the lords succeeded in increasing their wealth by means of improving their ability coercively to redistribute income away from the peasantry, they further limited the agricultural economy's capacity to improve. Increased rents in whatever form reduced the peasants' ability to make investments in the means of production. Meanwhile, the lords' allocation of their income to military followers and equipment and to luxury consumption ensured that the social surplus was used unproductively, indeed wasted. To the extent – more or less – that the lords increased their income, the agricultural economy was undermined.

(ii) *Population growth* (Postan 1966)

The long-term tendency to the decline of agricultural productivity thus conditioned by the feudal structure of property was realized in practice as a consequence of rising population. The peasants' possession of land allowed children to accede to plots and, on that basis, to form families at a relatively early age. Married couples, as noted, had an incentive to have many children, both to provide insurance for their old age and to assure that the line would be continued. The result was that all across the European feudal economy we witness a powerful tendency to population growth from around the beginning of the 12th century, which led, almost everywhere, to a doubling of population over the following of two centuries.

(iii) *Colonization* (Postan 1966; Duby 1968)

The only significant method by which the feudal economy achieved real growth and counteracted the tendency to declining agricultural productivity was by way of opening up new land for cultivation. Indeed, economic development in feudal Europe may be understood, at one level, in terms of the familiar race between the growth of the area of settlement and the growth of population. During the 12th and 13th centuries, feudal Europe was the scene of great movements of colonization, as settlers pushed eastward across the Elbe and southward into Spain, while reclaiming portions of the North Sea in what became the Netherlands. The opening of new land did, for a time, counteract and delay the decline of agricultural productivity. Nevertheless, in the long run – as expansion continued, as less fertile land was brought into cultivation, and as the man/land ratio rose – rents rose, food prices increased, and the terms of trade increasingly favoured agricultural as opposed to industrial goods. At various points during the 13th and early 14th centuries, all across Europe, population and production appear to have reached their upper limits, and there began to ensue a process of demographic adjustment along Malthusian lines.

(iv) *Political accumulation or state building* (Dobb 1946; Anderson 1974; Brenner 1982)

Given the limited potential for developing the agricultural productive forces and the limited supply of cultivable land, the lordly class, as noted, tended to find the buildup of the means of force for the purpose of redistributing income to be the best route for amassing wealth. Indeed, the lords found themselves more or less *obliged* to try to increase their income in order to finance the build-up of their capacity to exert politico-military power. This was, first of all, because they could not easily escape the politico-military conflict or competition that was the inevitable consequence of the individual lords' direct possession of the means of

force (the indispensable requirement for their maintenance as members of the ruling class over and against the peasants) and thus of the wide dispersal of the means of coercion throughout the society. It was, secondly, because they had to confront increasingly well-organized peasant communities and, as feudal society expanded geographically, to counteract the effects of increasing peasant mobility.

In the first instance, of course, military-politico efficacy required the collecting and organizing of followers. But to gain and retain the loyalty of their followers the overlords had to feed and equip them and, in the long run, competitively reward them. Minimally, the overlord's household had to become a focus of lavish display, conspicuous consumption and gift-giving, on par with that of other overlords. But beyond this, it was generally necessary to provide followers with the means to maintain their status as members of the dominant class – that is, a permanent source of income, requiring a grant of land with associated lordly prerogatives (classically the fief). But naturally such grants tended to increase the followers' independence from the overlords, leading to renewed potential for disorganization, fragmentation and anarchy. This was the perennial problem of all forms of patrimonial rule and at the centre of feudal concerns from the beginning. The tendency to fragmentation was, moreover, exacerbated as a result of the pressure to divide lordships and lands among children. To an important degree, then, feudal evolution may be understood as a product of lordly efforts to counteract political fragmentation and to construct firmer intra-lordly bonds with the purpose of withstanding intra-lordly politico-military competition and indeed of carrying on the successful warfare that provided the best means to amass the wealth ultimately required to maintain feudal solidarity. This meant not only the development of better weapons and improved military organization, but also the creation of larger and more sophisticated political institutions, and naturally entailed increased military and luxury consumption.

Actually to achieve more effective political organization of lordly groups required political innovation. Speaking broadly, the constitution of military bands around a leading warlord for external warfare, especially conquest, most often provided the initial basis of intra-lordly cohesion. This served as the foundation for developing more effective collaboration within the group of lords for the protection of one another's property and for controlling the peasantry. As a further step in this direction, the overlord would establish his pre-eminence in settling disputes among his vassals (as in Norman England). Next, the leading lord might extend feudal centralization by establishing immediate relations with the undertenants of his vassals. One way this took place was through constructing direct ties of dependence with these rear vassals (as in 11th-century England). More generally, it was accomplished by the extension of central justice to ever broader layers of the lordly class, indeed the free population as a whole. Sometimes the growth of central justice was achieved through the more or less conscious collaboration of the aristocracy as a whole (as in 12th-century England). On other occasions it had to be accomplished through more conflicted processes whereby the leading lord (monarch, prince) would accept appeals over the heads of his vassals from their courts (as in medieval France). Ultimately, the feudal state could be further strengthened only by the levying of taxes, and this almost always required the constitution of representative assemblies of the lordly class.

This is not to say that a high level of lordly organization was always required. Nor is it to argue that state building took place as an automatic or universal process. At the frontiers of European feudal society, to the south and east, colonization long remained an easy option, and there was relatively little (internally generated) pressure upon the lordly class to improve its self-organization. At the same time, just because stronger feudal states might become necessary did not always determine that they could be successfully constructed. Witness the failure of the German kings to strengthen their feudal state in the 12th century, and the long-term strengthening of the German principalities which ensued. The point is that, to the degree that disorganization and competition prevailed within and between

groups of feudal lords, they would tend to be that much more vulnerable not only to depredations from the outside, but to the erosion of their very dominance over the peasants. The French feudal aristocracy thus paid a heavy price for their early, highly decentralized feudal organization, suffering not only significant losses of territory to the Anglo-Normans, but a serious reduction in their control over peasant communities and a consequent decline in dues. The French aristocracy's later recovery and successes may be attributed, at least in large part, to their evolution of a new, more centralized, more tightly knit form of political organization – the tax/office state, where property in office (rather than lordship/land) gave the aristocracy rights to a share in centralized taxation (rather than feudal rent) from the peasants. In sum, the economic success of individual lords, or groups of them, does seem to have depended upon successful feudal state building, and the long-term trend throughout Europe, from the 11th through to the 17th century, appears to have been towards ever more powerful and sophisticated feudal states.

Trade, Towns and Feudal Crisis

The growing requirements of the lordly class for the weaponry and luxury goods (especially fine textiles) needed to carry on intra-feudal politico-military competition were at the source of the expansion of commerce in feudal Europe. The growth of trade made possible the rise of a circuit of interdependent productions in which the artisan-produced manufactures of the towns were exchanged for peasant-produced necessities (food) and raw materials, appropriated by the lords and sold to merchant middlemen. Great towns thus emerged in Flanders and north Italy in the 11th and 12th centuries on the basis of their industries' ability to capture a preponderance of the demand for textiles and armaments of the European lordly class as a whole.

In the first instance, the growth of this social division of labour within feudal society benefited the lords, for it reduced costs through increasing specialization, thus making luxury goods relatively cheaper. Nevertheless, in the long run it meant a growing disproportion between productive and unproductive labour in the economy as a whole, for little of the output of the growing urban centres went back into production to augment the means of production or the means of subsistence of the direct peasant producers; it went instead to military destruction and conspicuous waste. Over time, increasingly sophisticated political structures and technically more advanced weaponry meant growing costs and thus increased unproductive expenditures. At the very time, then, that the agricultural economy was reaching its limits, the weight of urban society upon it grew significantly, inviting serious disruption.

Because the growth of lordly consumption proceeded in response to the requirements of intra-feudal competition in an era of increasingly well-constructed feudal states, the lords could not take into account its effect on the underlying agricultural productive structure. All else being equal, the growth of population beyond the resources to feed it could have been expected to call forth a Malthusian adjustment, and most of Europe did witness the onset of famine and the beginning of demographic downturn in the early 14th century. Nevertheless, while the decline of population meant fewer mouths to feed with the available resources, it also meant fewer rent-paying tenants and so, in general, lower returns to the lords. The decline in seigneurial incomes induced the lords to seek to increase their demands on the peasantry, as well as to initiate military attacks upon one another. The peasants were thus subjected to increasing rents and the ravages of warfare at the very moment that their capacity to respond was at its weakest, and their ability to produce and to feed themselves was further undermined. Further population decline brought further reductions in revenue leading to further lordly demands – resulting in a downward spiral which was not reversed in many places for more than a century. The lordly revenue crisis and the ensuing seigneurial reaction thus prevented the normal Malthusian return to equilibrium. A general socio-economic crisis, the product of the overall feudal class/political system, rather than a mere

Malthusian downturn, gripped the European agrarian economy until the middle of the 15th century (Dobb 1946; Hilton 1969; Bois 1976; Brenner 1982).

In the long run, feudal crisis brought its own solution. With the decline of population, peasant cultivation drew back onto the better land, making for the potential of increased output per capita and growing peasant surpluses. Meanwhile, civil and external warfare seem to have abated, a reflection perhaps of the exhaustion of the lordly class, and the weight of ruling class exactions on the peasantry declined correspondingly, especially as the peasants were now in a far better position to pay. The upshot was a new period of population increase and expansion of the area under cultivation, of the growth of European commerce, industry and towns, and, ultimately, of the familiar outrunning of production by population. Meanwhile, lordly political organization continued to improve, feudal states continued to grow, intra-feudal competition continued to intensify, and, over the long run, lordly demands on the peasants continued to increase even as the capacity of the peasantry began, once again, to decline. By the end of the 16th century one witnesses, through most of Europe, a descent into the 'general crisis of the 17th century' which took a form very similar to that of the 'general crisis of the 14th and 15th centuries'. Clearly, through most of Europe, the old feudal property relations persisted, undergirding the repetition of established patterns of feudal economic non-development.

Approaches to Transition

It is an implication of the foregoing analysis that so long as feudal property relations persisted, the repetition of the same long-term economic patterns could be expected. So long as feudal property relations obtained, lords and peasants could be expected to find it rational to adopt the same patterns of individual economic behaviour; in consequence, one could expect the same long-term cyclical tendencies to declining agricultural productivity, population growth, and the opening of new land, issuing in a tendency to Malthusian adjustment but overlaid by a continuation of the secular tendency to lordly state building and growing unproductive expenditures. Generally speaking, so long as feudal property relations obtained, no inauguration of a long-term pattern of modern economic growth could be expected. From these premises, it is logical to conclude that the onset of economic development depended on the transformation of feudal property relations into capitalist property relations, and that indeed is the point of departure of a long line of theorists and historians (Marx 1894; Dobb 1946; Hilton 1969; Bois 1976).

Nevertheless, beginning with Adam Smith himself, a whole school of historically sensitive theorists have found it quite possible to ignore, or sharply to downplay, the problem of the transformation of property relations and of social relationships more generally in seeking to explain economic development. These theorists naturally refuse to go along with the Adam Smith of *Wealth of Nations* Book I in contending that the mere application of individual economic rationality will, directly and automatically, bring economic development. They nevertheless follow the Adam Smith of *Wealth of Nations* Book III in arguing that, given the appearance of certain specific, *quite-reasonable-to-expect* exogenous economic stimuli, rational self-interested individuals can indeed be expected to take economic actions which will detonate a pattern of modern economic growth. Specifically, it is their hypothesis that the growth of commerce, an enormously widespread if not universal phenomenon of human societies, systematically has led pre-capitalist economic actors to assume capitalist motivations or goals, to adopt capitalist norms of economic behaviour, and, eventually, to bring about the transformation of pre-capitalist to capitalist property relations. It is undoubtedly because Adam Smith and his followers have believed that the growth of exchange will *in itself* sooner or later create the necessary conditions for modern economic growth that they have not greatly concerned themselves with these conditions or viewed their emergence as a problem which needs addressing.

Thus, Smith and a long line of followers, prominently including the economic historian of

medieval Europe Henri Pirenne and the Marxist economist Paul Sweezy, have all produced analyses which follow essentially the same progression. First, merchants, emanating from outside feudal society, offer previously unobtainable products to lords and peasants who hitherto had produced only for subsistence. This is understood as a more or less epoch-making historical event, an original rise of trade. Next, the very opportunity to purchase these new commodities induces the individual economic actors to adopt business-like attitudes and capitalist motivations, specifically to relinquish their norm of production for subsistence and to adopt the economic strategy of capitalists-in-embryo – viz., production for exchange so as to maximize returns by way of cost cutting. Third, since pre-capitalist property relations, marked by the producers' possession of the means of subsistence and by the lord's extraction of a surplus by means of extra-economic coercion, prevent the individual economic actors from most effectively deploying their resources to maximize exchange values, both lords and peasants move, on a unit-by-unit basis, to transform these property relations in the direction of capitalist property relations. In particular, the lords dispense with their (unproductive) military followers and military luxury expenditures; they free their hitherto dominated peasant producers; they expropriate these peasants from the land; then, finally, they enter into contractual relations with these free, expropriated peasants. This gives rise, within each unit to the installation of free, necessarily commercialized (market dependent) tenants on economic leases, who, ultimately, hire wage labourers. The end result is the establishment of capitalist property relations and capitalist economic norms in the society as a whole and the onset of economic development (Smith 1776; Pirenne 1937; Sweezy 1950).

The foregoing argument of what might be called the Smithian school is designed, implicitly or explicitly, to show how the rise of exchange in a feudal setting, in itself creates the conditions under which rational economic actors will pursue self-interested action which leads, on an economy-wide basis, to modern economic growth. Nevertheless, the validity of each step in the Smithian argument can be, and has been, challenged by those who take as their point of departure the historically established property relations. It is the essence of their position that the Smithians can sustain their argument only by failing sufficiently to understand what patterns of economic activity individual lords and peasants will find it rational to adopt in response to the rise of trade, *given* the prevalence of feudal property relations (Marx 1894; Dobb 1946; Bois 1976).

In the first place, although long-distance merchants may bring to feudal lords and peasants commodities they could not previously obtain, the merchants' mere offer of these commodities cannot ensure that the lords and peasants will, in turn, put their own products on the market in order to buy them. Given the existence of feudal property relations, both lords and peasants may be assumed to have everything they need to maintain themselves. The opportunity to buy new goods may very well make it possible for the pre-capitalist economic actors to increase or enrich their consumption, but this does not mean that they will take advantage of this opportunity. The increased potential for exchange simply cannot determine that exchange will increase (Luxemburg 1913).

Secondly, even where the appearance of new goods brought by merchants does induce the lords to try to increase their consumption by raising their output and increasing the degree to which they orient their production towards exchange, this will hardly lead them to find it in their rational self-interest to dismantle, in piecemeal fashion, the existing feudal property relations by freeing and expropriating their peasants. Given the reproduction of feudal property relations by communities of feudal lords and peasants, the individual lords can hardly find it in their rational self-interests to free their peasants, for they would lose thereby their very ability to exploit them, and thus their ability to make an income. The point is that, once freed from the lord's extra-economic domination, his *possessing* peasants would have no need to pay *any* levy to him, let alone increase the quality and quantity of their work for him. Moreover, even if the lord could, at one and the same time, free *and*

expropriate his peasants, he would still lose by the resulting transformation of his unfree peasant possessors into free landless tenants and wage labourers, for the newly landless tenants or wage labourers would have no reason to stay and work for their former lord or to take up a lease from him.

To the degree, then, that lords sought to increase their output in response to trade, they appear to have found it in their rational self-interest not to transform but to intensify the pre-capitalist property relations. Because they found it, on the one hand, difficult to get their possessing peasants effectively to use more productive techniques on their estates, and, on the other hand, irrational to instal capitalist property relations within their units, they seem to have had little choice but to try to do so within the constraints imposed by feudal property relations – by increasing their levies on the direct producers in money, kind or labour. To make this possible, they had no choice but to try to strengthen their institutionalized relationship of domination over their peasants, by investing in improved means of coercion and by improving the politico-military organization of their lordly groups. It needs to be emphasized that the lords could not be sure they could succeed in this, for the peasants would likely resist, and perhaps successfully. But in so far as the lords could dictate terms, this was the route they found most promising. Witness the growth of demesne farming in response the growth of the London market in 13th-century England or, more spectacularly, the rise of a neo-serfdom throughout later medieval and early modern eastern Europe in response to the growth of trade with the west (Dobb 1946).

Finally, it needs to be noted that the sorts of products on the market which were most likely to stimulate the exploiters to try to increase their income for the purpose of trade were goods which 'fit' their specific reproductive needs. These were not producer goods but, on the contrary, means of consumption – specifically, materials useful for building up the exploiters' political and military strength. They were certainly not luxury goods in the ordinary sense of superfluities, for they were, in fact, necessities for the exploiters. But they were luxuries in that their production involved a subtraction from the means available to the economy to expand its fundamental productive base.

Paradoxically, then, to the extent that the rise of trading opportunities, *in itself,* can be expected to affect precapitalist economies, it is likely to bring about not the loosening but the tightening of pre-capitalist property forms, the growth of unproductive expenditure, and the quickening not of economic growth but of stagnation and decline.

From Feudalism to Capitalism

The onset of modern economic growth thus appears to have required the break-up of pre-capitalist property relations characterized by the peasants' possession of their means of subsistence and the lords' surplus extraction by extra-economic compulsion. Nevertheless, neither the regular recurrence of system-wide socio-economic crisis nor the widespread growth of exchange could, in themselves, accomplish this. The problem which thus emerges is how feudal property relations could ever have been transformed.

To begin to confront this question, one can advance two basic hypotheses which follow more or less directly from the central themes of this article:

1. In so far as lords and peasants, acting either individually or as organized into communities, were able to realize their conscious goals, they succeeded, in one way or another, in maintaining pre-capitalist property forms. This is to say, once again, that the patterns of economic activity that individual lords and peasants found it reasonable to pursue could not aim at transforming the feudal property structure. It is also to emphasize that, because peasants and lords organized themselves into communities for the very purpose of maintaining and strengthening, respectively, peasant possession and the institutionalized relationships required for taking a feudal rent by extra-economic coercion, lords and peasants acting as communities were unlikely to aim at undermining feudal property forms. Peasants might, through

collective action, conceivably have reduced to zero the lords' levies and eliminated the lords' domination; but, even in this extreme case, they would have ended up constituting a community of peasants fully in possession of their means of subsistence, with all of the barriers to economic development entailed by that set of property relations. Were the lords, on the other hand, to have succeeded to the greatest extent conceivable in overcoming peasant resistance, they would only to that degree have strengthened their controls over the peasants and increased their rate of rent, thus tightening feudal property relations.

2. Where breakthroughs took place to modern economic growth in later medieval and early modern Europe, these must be understood as *unintended consequences* of the actions by individual lords and peasants and by lordly communities and peasant communities in seeking to maintain themselves as lords and peasants in feudal ways. In other words, the initial transitions from feudal to capitalist property relations resulted from the attempts by feudal economic actors, as individuals and collectivities, to follow feudal economic norms or to reproduce feudal property relations under conditions where, doing so, actually had the effect – for various reasons – of undermining those relations.

To give substance to these hypotheses would require a lengthy historical discussion. It is here possible only to note a broad contrast in the historical evolutions of the different European regions during the late medieval and early modern periods. Through most of pre-industrial Europe, east and west, varying processes of class formation brought, in one form or another, the reproduction of feudal property relations and, in turn, the repetition of long-term developmental patterns familiar from the medieval period. However, in a few European regions, feudal property relations dissolved themselves, giving rise, for the first time, to essentially modern processes of economic development.

Thus, through much of later medieval and early modern western Europe (France and parts of western Germany), although peasants succeeded in very much strengthening peasant possession, winning their freedom and destroying all forms of surplus extraction by extra-economic coercion by individual lords, the lords succeeded, in response, in maintaining themselves by means of constituting a new, more potent form of now-collective surplus extraction by extra-economic compulsion, the tax/office state. At the same time, throughout late medieval and early modern eastern Europe, despite the peasants' initially very powerful rights in the land and the lords' initially very weak feudal controls, the lords ended up erecting an extremely tight form of individual lordly domination and surplus extraction by extra-economic compulsion – serf-operated demesne production. The consequence of these reconsolidations of essentially feudal property relations throughout most of Europe, east and west, was the reappearance throughout most of Europe during the early modern period of the same trends towards demographically powered expansion, towards the continued build-up of larger and more sophisticated states and, ultimately, towards socio-economic crisis as had characterized the medieval period.

The evolution of property relations in late medieval and early modern England was in some contrast to that of both eastern and (most of) western Europe, with epochal consequences for the long-term pattern of economic development. During this period, English lords, unlike those in eastern Europe, failed, as did those throughout almost all of western Europe, in their attempts to maintain, let alone intensify, their extra-economic controls over their peasantry. On the other hand, the English lords, unlike those throughout much of western Europe, did ultimately succeed in maintaining their positions by means of preventing their customary tenants from achieving full property in their plots. They were able, in consequence, to consign these tenants to leasehold status, and thus to assert their own full property in the land.

The unintended consequence of the actions of English peasants and lords aiming to maintain themselves as peasants and lords in feudal ways was thus to introduce a new system of now-capitalist property relations in which the direct producers were free from the lords' extra-economic domination but also separated from their full means of reproduction (subsistence).

In the upshot, tenants without direct access to their means of reproduction had no choice but to produce competitively for exchange and thus, so far as possible, to specialize, accumulate and innovate. At the same time, the landlords found themselves obliged to create larger, consolidated and well-equipped farms if they wished to attract the most productive tenants. The long-run results were epoch making. Under the pressures of competition, processes of differentiation led to the emergence of an entrepreneurial class of capitalist tenant farmers who were ultimately able to employ wage labourers. Meanwhile, the drive to cut costs in agricultural production ultimately brought about an agricultural revolution, as market-dependent farmers were obliged to adopt techniques which long had been available, but long eschewed by possessing peasants who would not intentionally take the risks of specialization, let alone make the necessary capital investments. The secular decline in food costs and the secular rise in living standards which resulted underpinned the movement of population off the land and into industry and made possible the rise of the home market. Industry and agriculture, for the first time, proved mutually supporting, rather than mutually competitive, and population increase served to stimulate economic growth rather than to undermine it. England experienced unbroken industrial and demographic growth right through the 17th and 18th centuries, which ultimately issued in the Industrial Revolution.

See Also

- ▶ Dobb, Maurice Herbert (1900–1976)
- ▶ Peasants
- ▶ Sweezy, Paul Marlor (1910–2004)

Bibliography

Anderson, P. 1974. *Passages from antiquity to feudalism*. London: New Left Books.

Bois, G. 1976. *La crise du féodalisme*. Paris: Editions EHESS.

Bloch, M. 1939–40. *Feudal society.* Trans. L.A. Manyon. Chicago: University of Chicago Press, 1961.

Brenner, R. 1982. The agrarian roots of European feudalism. In *The Brenner debate: Agrarian class structure and economic development in preindustrial Europe*, ed. T.H. Aston. Cambridge: Cambridge University Press. 1985.

Dobb, M. 1946. Studies in the development of capitalism. London/New York: Routledge & Kegan Paul/International Publishers, 1947.

Duby, G. 1968. *Rural economy and country life in the medieval West.* Trans. C. Postan. Columbia: University of South Carolina Press.

Duby, G. 1973. *The early growth of the European economy.* Trans. H.B. Clarke. Ithaca: Cornell University Press, 1974.

Duby, G. 1978. *The three orders of society.* Trans. T.N. Bisson. Chicago: University of Chicago Press, 1980.

Ganshof, F.L. 1947. *Feudalism.* Trans. P. Grierson. New York: Harper & Row, 1961.

Hilton, R.H. 1966. *A medieval society.* New York: Wiley.

Hilton, R.H. 1969. *The decline of serfdom*. London: Macmillan.

Luxemburg, R. 1913. *The accumulation of capital.* Trans. A. Schwarzschild. New York: .Monthly Review Press, 1968.

Marx, K. 1894. *Capital, Vol. 3*. New York: International Publishers, 1967.

Pirenne, H. 1937. *Economic and social history of medieval Europe*. New York: Harcourt Brace & Co.

Postan, M.M. 1966. Medieval agrarian society in its prime: England. In *The Cambridge economic history of Europe, vol. 1: The agrarian life of the middle ages*, ed. M.M. Postan and H.J. Habakkuk, 2nd ed. Cambridge: Cambridge University Press.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of Nations,* ed. R.H. Campbell, A.S. Skinner and W.B. Todd. Oxford: Clarendon Press, 1976.

Strayer, J.R. 1965. *Feudalism*. New York: Van Nostrand Reinhold.

Sweezy, P. 1950. The transition from feudalism to capitalism. *Science and Society* 14 (2): 134–157.

Weber, M. 1956. Patriarchalism and patrimonialism. Feudalism, Standestaat, and patrimonialism. In *Economy and society*, 2 vols, ed. G. Roth and C. Wittich. Berkeley: University of California Press, 1978.

Fiat Money

Neil Wallace

Abstract

Fiat money is an intrinsically useless object that serves as a medium of exchange. One challenge

is to construct models that depict the ancient notion that a medium of exchange is beneficial. Another is to construct models in which the medium of exchange has a low rate of return. This article reviews how those challenges have been approached and argues that progress has been achieved by taking seriously some old ideas about the circumstances in which money is helpful and about the desirable properties of money: money is helpful when there are absence-of-double-coincidence difficulties that cannot be easily overcome with credit; and a good money has desirable physical properties – recognizability, portability and divisibility.

Keywords

Absence of doublecoincidence; Arrow–Debreu model; Asymmetric information; Cash-in-advance models; Central banking; Commitment; Commodity money; Cournot quantity game; Credit; Fiat money; Friedman rule; Imperfect monitoring; Incentive feasibility; Incomplete markets; Infinite horizons; Inside and outside money; Large-family models; Medium of exchange; Money; Nash equilibrium; Open-market operations; Perfect monitoring; Pooling equilibria; Production functions; Quantity theory of money; Risk sharing; Shapley–Shubik trading posts; Walras's Law

JEL Classifications

N0

An object is often said to qualify as *money* if it plays one or more of the following roles: a unit of account, a medium of exchange, a store of value. The first and third seem insufficient. The Arrow–Debreu model with prices expressed in terms of either an abstract numeraire or one of the goods is not a model of a monetary economy. Neither is every model that contains an asset or durable good. That leaves the medium-of-exchange function: an object is a medium of exchange if it appears in many transactions – in the sense of a Clower (1967) transaction matrix.

As regards kinds of money, one distinction is between outside money, such as gold coins, and inside (private sector) money, such as demand deposits. (The quantity of outside money is unaffected by consolidation over the balance sheets of everyone in the economy, while the quantity of inside money disappears when that consolidation is performed – an inside money being someone's asset and someone else's liability.) Among outside monies, a distinction is usually made between commodity and fiat money. A commodity money is an object that has intrinsic value as a consumption good or as an input, while a fiat money does not.

One challenge is to construct models that depict the ancient notion that a medium of exchange is beneficial. (This notion goes back at least to the Roman jurist Paulus who said: 'Since occasions where two persons can just satisfy each other's desires are rarely met, a material was chosen to serve as a general medium of exchange' –Monroe 1966.) Another is to construct models in which media of exchange are relatively poor stores of value, have low rates of return. And accompanying those challenges is a wide range of related policy questions. How, if at all, should inside money be regulated? How should a government monopoly on outside money be managed? Should there be country-specific outside monies?

Progress in meeting those challenges and in addressing policy questions has come about by taking seriously some old ideas: money is helpful when there are absence-of-double-coincidence difficulties that cannot be easily overcome with credit; and a good money has some desirable physical properties – recognizability, portability, and divisibility. In order to better appreciate the challenges and the progress, it is helpful to review the history of monetary theory.

The Classical Dichotomy

At the beginning of the 20th century, the dominant economic theory was a two-part model: a rudimentary Arrow–Debreu theory of relative prices and allocations; and a quantity-theory equation that was often interpreted as a supply-equals-demand for money equation. As was

widely recognized, this model suffers from a blatant inconsistency. Everybody in the model is completely described in the theory of relative prices and allocations. Who, then, holds money, which is not one of goods in the relative price-allocation part of model? Patinkin (1951) called attention to this inconsistency by pointing out that the model fails to satisfy Walras's Law.

The model has other defects. Because the model does not describe transactions, it is silent about whether money is a medium of exchange. Whether it is or not, money is not helpful in the model because allocations are determined exactly as they would be in its absence. And, as was widely recognized, the real return on money in the model – determined entirely by the time path of the stock of money and its effect on the time path of the price level – could be less than, equal to, or greater than the real interest rate determined in the relative price part of the model. The third possibility was viewed as problematic because people would then, presumably, hold only money.

Notice, by the way, that money in the above model is implicitly fiat money and that holdings of it are minimized subject to being able to carry out transactions. Neither was an obvious feature of the economies to which the theory was applied for centuries. For most of that time, money was in fact a commodity and one that may not have been a poor store of value – if only because few alternatives were available. The distinction between commodity and fiat money may not be important because for some specifications of the intrinsic value of commodity money, the value of commodity money is determined in the same way as the value of fiat money (see, for example, Samuelson 1968; Sargent and Wallace 1983). The implicit assumption that money is a poor store of value is more significant because it means that money cannot be treated as an ordinary asset.

Real Balances in Utility or Production Functions

The first models to overcome the blatant inconsistency of the classical dichotomy and to, in some way, integrate value and monetary theory were models of fiat money in which its quantity and its price were arguments of utility or production functions (see Samuelson 1961). Such models are consistent with individual endowments of money and have equilibria in which it has value.

The models were intended to overcome the inconsistency of the classical dichotomy, while preserving as much of the relative price part of the model as possible. However, not everything was preserved. After explaining why real balances, not nominal balances, are introduced as an additional argument of utility functions, Samuelson (1961, p. 119) says, 'This is not the only case in which economists have found it necessary to introduce prices into the indifference loci; there is also the example of goods which have snob appeal, or scarcity appeal. . .' Samuelson (1968) describes the welfare consequences of his formulation: the failure of the first welfare theorem. That failure should not be surprising; putting prices into utility or production functions is a back-door way of introducing externalities. The failure gave rise to the vast literature on the so-called Friedman rule: tax to support the payment of interest on money either explicitly or through deflation.

A desirable feature of these models is that money cannot have a higher pecuniary real return than other assets. The models treat real balances like clothing or refrigerators. Such assets throw off services and, therefore, in equilibrium have lower pecuniary rates of return than assets like bonds that do not throw off services.

Cash in Advance and Trading Posts

Utility or production functions with real balances as arguments were always regarded as *indirect* functions. If so, then there ought to be a direct or underlying model. One suggestion for the underlying model is a model in which the Arrow–Debreu budget set is replaced by separate sets which insure that money will appear in many trades (see Clower 1967). Some goods can be purchased only with money and the sellers of those goods who receive money can use that money only in subsequent trades. Such models,

dubbed cash-in-advance models, are special cases of models of *incomplete markets* (see, for example, Magill and Quinzii 2006).

Viewed that way, cash-in-advance models depart from the Arrow–Debreu model by amending its equilibrium concept. Shubik (1973) adopts that way of modelling money, but insists that trade be modelled as an explicit game. In particular, he suggests that it be modelled using what are called Shapley–Shubik trading posts, with each post defined by the pair of objects traded at the post. In static versions of that model in which the game is modelled as the simultaneous choice of quantities (a version of a Cournot quantity game), inactivity at any subset of posts (including all posts) is a Nash equilibrium. Such inactivity has been used as a rationale for selecting a subset of posts that produces the kind of transaction matrix we observe – for example, some goods cannot be traded for anything other than money and, in a multi-country context, some goods can be traded only for home money. However, Krishna (2005) questions the robustness of shutting down posts in which goods trade for assets that dominate money in rate of return.

Starr and Stinchcombe (1999) use a version of this model with fixed costs of operating a post to suggest that scale economies can imply that the efficient arrangement of posts when there are $n + 1$ objects, n goods and money is a monetary structure: at each of n active posts, money trades for one of the goods. Howitt (2005) uses an infinite-horizon version of that model with utility-maximizing agents who operate the posts to argue that there can be equilibria with that monetary structure of posts.

Imperfect Monitoring and Money

A different approach to modelling money is to depart from the environment of the Arrow–Debreu model – in particular, from its assumptions about commitment and information. Implicit in the absence-of-double-coincidence rationale for money is that the two persons cannot commit to future actions and are strangers. After all, a student in a class is more likely to say to a neighbouring student 'lend me a pencil' than 'sell me a pencil'. More generally, in order that absence of double coincidence be a basis for a beneficial role for money, it must be augmented by no-commitment and by informational assumptions that inhibit the use of credit in its most general sense–informational assumptions that in game theory are called *imperfect monitoring.*

One of the first discussions of the informational assumptions is in Ostroy (1973). Townsend (1989) uses imperfect monitoring in an explicit intertemporal model and Kocherlakota (1998) further formalizes it. This work treats fiat money as a mechanism whose only role is to provide evidence of previous actions that would otherwise not be known. Fiat money, a physical object, can play that role because, counterfeiting aside, others can say 'show me' if one tries to overstate ones holdings of it.

The potentially crucial role of imperfect monitoring can be illustrated by considering the well-known risk-sharing model in Green (1987) and the variant of it studied by Levine (1990). There is a non-atomic measure of people who have identical preferences and maximize expected discounted utility. The model is one of pure exchange with a single good at each date. At each date, each person receives an endowment realization from a two-point set (high or low), where realizations are i.i.d. among people at a date and over time and are private information. Green studies a version of this model with perfect monitoring: at each date, each person makes a report about the person's endowment realization, a report which in the future is associated with that person.

Levine (1990) studies a variant of this model, but assumes no monitoring at all. In his version, no announcement or action made by a person at a date is associated with that person in the future. Moreover, if endowments are treated as owned by individuals, then under Levine's assumption, there is a role for money even if endowment realizations are public information. If there is no way to remember in the future that a person with a high endowment surrendered some of it, then the person will not surrender it – except for something that the person can carry into the future. In a pure-exchange setting, that thing can only be fiat money.

Pairwise Meetings

Absence of double coincidence is almost always described in terms of meetings between two people. That, of course, is very different from having everyone together or at least connected as in the Arrow–Debreu model. But, if the role of such pairwise meetings is only to prevent quid pro quo trade in commodities, then it is unnecessary. Such trade cannot happen in Green (1987), even in deterministic versions of it. So why bother with models of pairwise meetings?

One reason is that Paulus and others were reporting what they were seeing: namely, exchanges between two people. Another reason to study such models is to investigate their implications for transactions. Kiyotaki and Wright (1989) are the first to succeed in formulating and analysing such a model. In a world with many objects, they study the relationship between the intrinsic storage properties of objects – in particular, the (utility) cost of storing them – and their role in exchange. In order to make headway on that question, they adopt simplifying assumptions: objects are indivisible, each person can hold at most one unit of some object, and the intrinsic storage quality of an object is modelled as a utility cost which once realized does not become part of the state of the economy. Even with those simplifying assumptions, their model is not simple because the state of the economy is a distribution of holdings of the different objects. Nevertheless, they could show that there can be steady states in which objects other than the least costly-to-store object can play a medium-of-exchange role. (For the welfare properties of different equilibria in their model, see Renero 1999.)

Still another reason for studying models with pairwise meetings is that such meetings can provide a rationale for imperfect monitoring. In a large economy, if people meet in pairs and, therefore, know only what they have experienced or what they have been told by people they meet, then imperfect monitoring emerges as an implication. This point of view is explored in non-monetary models in Kandori (1992) and in monetary models in Kocherlakota (1998) and Araujo (2004). Finally, models of pairwise meetings are attractive settings for exploring the consequences of imperfect recognizability and imperfect divisibility of money and other assets.

Models of pairwise meetings, however, also come with complications. One is the wide range of equilibrium concepts used to answer the old question: what do a pair who meet to trade do? One approach taken in the literature is descriptive – for example, the buyer and the seller make alternating offers, buyers make take-it-or-leave-it offers, or sellers commit to posted prices. Another approach explores all implementable outcomes subject either to individual defection or to such defection and cooperative defection by the pair in the meeting.

Another complication is the endogeneity of the distribution of assets. Such endogeneity also arises in models in which fiat money is the only durable object, in which people can hold more than one unit of money, and in which the meeting process gives rise to a distribution of outcomes – a person can end up buying, selling, or not trading. Obviously, in such models we do not expect to obtain simple closed-from solutions for equilibria or even steady states.

One response is to accept the endogeneity and to derive results for the model despite not having closed-form solutions (see Green and Zhou 1998; Molico 2006; Zhu 2003, 2005). Another is to avoid it: by using the so-called *large-family* model (see Shi 1997); by using a setting in which pairwise meetings alternate in some fashion with centralized meetings in which preferences are quasi-linear (see Lagos and Wright 2005); or by using some other meeting process that lends itself to a simple or degenerate distribution of money.

Applications

New theoretical work should provide insights previously unavailable – insights about seemingly paradoxical observations or policies or both.

Outside Money, Credit and Cashless Economies

If we maintain the innocuous assumption that people cannot commit to future actions, then a

model economy with perfect monitoring has no role for money, while one with no monitoring has no role for credit. Therefore, in order to find roles for both money and credit, we should study models with some, but not perfect, monitoring.

Several alternative formulations of such imperfect monitoring have been studied. Kocherlakota and Wallace (1998) use the pairwise setting in Trejos and Wright (1995) and Shi (1995), and assume that there is a lag in updating the public record of individual actions. They show that the set of implementable allocations is larger the shorter the lag – an obvious result, but one that represents the sense in which technological improvements that allow better monitoring improve trade outcomes. Cavalcanti and Wallace (1999) use the same background model, but assume that some people are perfectly monitored and others not at all. They permit each person to issue perfectly recognizable durable objects that are specific to the person, objects that are best interpreted as transferable trade-credit instruments. They show that the set of implementable outcomes in which such instruments are not valued (or are prohibited) is a strict subset of those in which such instruments issued by monitored people are valued. (Kocherlakota 2002, shows that there is a way to support efficient allocations in such models using only spot trade with money. However, his punishment scheme would not survive allowing either the defector or the non-defector to move first in a meeting.) Aiyagari and Williamson (2000) use an environment that is close to that of Green (1987), but assume that a report to the planner can be made with some probability less than 1. Their focus is on how competitive trade in money influences what the planner can achieve.

Obviously, limiting cases of the above formulations of imperfect monitoring give rise to what can be interpreted as cashless economies. Although there are many conceptions of cashless economies, one of which is the Arrow–Debreu model, the above formulations have the desirable property that the cashless limit is a limit of a cash economy in which a medium of exchange plays a beneficial role. Moreover, because the cashless economy is achieved by taking a limit with respect to monitoring while maintaining the no-commitment assumption, the cashless limit is not an Arrow-Debreu model.

In Cavalcanti and Wallace (1999) and Cavalcanti et al. (1999), the money issued by monitored people is used by and passed around among nonmonitored people. Wallace and Zhu (2007) use that idea to offer a new interpretation of the paradox concerning banknote issue pointed out by Friedman and Schwartz (1963). Toward the end of 19th century, many countries permitted banks to issue payable-to-the-bearer notes subject to redemption on demand and to collateral restrictions. In the United States and, presumably, in other countries, those systems seemed to give rise to a failure of an arbitrage condition: the yields on eligible collateral often seemed too high to be reconciled with their use as collateral for note issue. Put differently, those systems seemed not to produce currencies that were elastic with respect to the yield on eligible collateral. The explanation offered by Wallace and Zhu has two components. First, the profitability of note issue depends on the implied float. Second, note issuers face a menu of opportunities for issuing notes – a menu that displays an inverse association between the magnitude of possible note placements and the implied float. The paradox results from treating the observed float as if it applied to all possible uses of notes, rather than taking into account the fact that high-placement low-float opportunities – for example, in organized financial markets – are not chosen. In Wallace and Zhu, the low-placement, high-float opportunities are in pairwise meetings.

Physical Properties of Assets

Discussions of money have often described desirable physical properties of media of exchange: recognizability, portability and divisibility. Implicit in any such discussion is the idea that those properties are scarce, are not shared equally by all objects. However, only recently have the consequences of such scarcity been explored.

Recognizability

Freeman (1985) and Williamson and Wright (1994) use imperfect recognizability of alternatives to fiat money to produce models in which fiat money is helpful. In Freeman, the alternative

to fiat money is a claim to long-lived capital. Under the assumption that such claims can be costlessly counterfeited, he argues that genuine claims cannot be traded competitively. Williamson and Wright use a model of pairwise matching *without* an absence-of-double-coincidence problem to show that imperfect recognizability of the (durable) goods is enough to make trade involving fiat money helpful.

In both of those models and many others, the holder of an asset knows more about it than at least some potential holders. (An exception is Huggett and Krasa 1996.) Models of pairwise meetings are attractive for studying the role of such imperfect recognizability because it is in such meetings, rather than in 'large markets', that asymmetric information about quality ought to be important. Moreover, if, as in Freeman or Williamson and Wright, the low-quality asset is worthless, then it gets traded when subject to asymmetric information only if it masquerades as being genuine – that is, only in a *pooling* equilibrium.

However, pooling equilibria do not always exist – at least if refinements on beliefs about off-equilibrium actions are imposed. It remains to be determined whether such refinements could be used to strengthen the Freeman result. In particular, could a small difference in counterfeiting costs between two assets – between fiat money and claims to capital, or home money and foreign money, or outside money and inside money – be enough to generate trade in one of the assets and no trade in the other even if the less-costly-to-counterfeit asset, as in Freeman, has a large rate-of-return advantage?

Portability

Townsend (1989) and Smith (2002) build models based on portability of fiat money and the lack of portability of capital. However, as they emphasize, the mere lack of portability of real capital needs to be supplemented by imperfect monitoring. And when supplemented by sufficiently imperfect monitoring, such models give rise to a role for fiat money that is very similar to its role in other absence-of-double-coincidence settings.

To see the similarity, consider a version of those models in which people meet in pairs and in which there is one good per date. When two people meet, suppose that they have available to them some amount of the good that can either be consumed or used as an input (investment) that will give output at the next date, but only at the same location. Moreover, suppose that one and only one of the two people will be at the same location at the next date. If there is no monitoring, then fiat money, despite having a lower real return than investment, can have a beneficial role – the same role it has in other absence-of-double-coincidence settings with no monitoring. That is, the stayer retains all the capital, while the leaver takes some fiat money. The absence of monitoring prevents the leaver from retaining a claim to any of the capital.

Divisibility

Historians of monetary systems and others have often noted that money was generally not available in conveniently small denominations (see, for example, Redish 2000; Sargent and Velde 2002). However, until recently no models described how such absence would inhibit trade. Models of pairwise meetings are an obvious candidate: if neither the buyer nor the seller has small change, then trade (even if lotteries are permitted) is inhibited. If the model is to have implications for optimal divisibility, then it should also contain something to limit divisibility. Lee et al. (2005) assume that there is a direct cost of carrying monetary items that is independent of denomination (that is, carrying thousands of pennies is very costly), while Lee and Wallace (2006) assume costs of producing and maintaining the stock of money that increase with divisibility.

Concluding Remarks

Why is it better to make assumptions about meeting patterns, information, and the physical characteristics of potential assets than about which markets are open or the pattern of transaction costs over objects? First, the former lends itself to standard notions of incentive feasibility, which is what we ought to mean by integrating monetary economics with the rest of economics. Second, such an approach meets the proof-of-the-pudding

criterion. Compare, for example, the results about inside money that can be obtained by working with the imperfect-monitoring point of view with what can be done with a cash-in-advance model.

But is such foundational work needed to deal with the nuts and bolts of monetary policy? It is generally agreed that open-market operations matter because the medium of exchange is a low-return asset and because the central bank has a monopoly on its supply. Can it be that beneficial management of that monopoly does not depend on how we explain the low return of the medium of exchange?

Finally, can we look forward to a monetary theory that in generality rivals the Arrow–Debreu model? Probably not. A need for a medium of exchange does not arise in every conceivable economy – think of Robinson Crusoe, even after he meets Friday, or of the Arrow–Debreu model. Such a need arises when there is some absence-of-double-coincidence difficulty that cannot be overcome with credit because people cannot commit to future actions and because there is imperfect monitoring. Those features may not lend themselves to a general formulation.

See Also

- ▶ Currency Competition
- ▶ Inside and Outside Money
- ▶ Money
- ▶ Money and General Equilibrium

Bibliography

Aiyagari, S., and S. Williamson. 2000. Money and dynamic credit arrangements with private information. *Journal of Economic Theory* 91: 248–279.

Araujo, L. 2004. Social norms and money. *Journal of Monetary Economics* 51: 241–256.

Cavalcanti, R., and N. Wallace. 1999. Inside and outside money as alternative media of exchange. *Journal of Money, Credit and Banking* 31: 443–457.

Cavalcanti, R., A. Erosa, and T. Temzelides. 1999. Private money and reserve management in a random matching model. *Journal of Political Economy* 107: 929–945.

Clower, R.W. 1967. A reconsideration of the microfoundations of monetary theory. *Western Economic Journal* 6: 1–8.

Freeman, S. 1985. Transaction costs and the optimal quantity of money. *Journal of Political Economy* 93: 146–157.

Friedman, M., and A. Schwartz. 1963. *A monetary history of the United States*. Princeton: Princeton University Press.

Green, E. 1987. Lending and the smoothing of uninsurable income. In *Contractual arrangements for intertemporal trade*, ed. E. Prescott and N. Wallace. Minneapolis: University of Minnesota Press.

Green, E., and R. Zhou. 1998. A rudimentary random-matching model with divisible money and prices. *Journal of Economic Theory* 81: 252–271.

Howitt, P. 2005. Beyond search: Fiat money in organized exchange. *International Economic Review* 46: 405–429.

Huggett, M., and S. Krasa. 1996. Money and storage in a differential information economy. *Economic Theory* 8: 191–210.

Kandori, M. 1992. Social norms and community enforcement. *Review of Economic Studies* 59: 63–80.

Kiyotaki, N., and R. Wright. 1989. On money as a medium of exchange. *Journal of Political Economy* 97: 927–954.

Kocherlakota, N. 1998. Money is memory. *Journal of Economic Theory* 81: 232–251.

Kocherlakota, N. 2002. The two-money theorem. *International Economic Review* 43: 333–346.

Kocherlakota, N., and N. Wallace. 1998. Optimal allocations with incomplete record-keeping and no-commitment. *Journal of Economic Theory* 81: 272–289.

Krishna, R.V. 2005. Non-robustness of the cash-in-advance equilibrium in the trading-post model. *Economics Bulletin* 5: 1–5.

Lagos, R., and R. Wright. 2005. A unified framework for monetary theory and policy analysis. *Journal of Political Economy* 113: 463–484.

Lee, M., and N. Wallace. 2006. Optimal divisibility of money when money is costly to produce. *Review of Economic Dynamics* 9: 541–556.

Lee, M., N. Wallace, and T. Zhu. 2005. Modeling denomination structures. *Econometrica* 73: 949–960.

Levine, D. 1990. Asset trading mechanisms and expansionary policy. *Journal of Economic Theory* 54: 148–164.

Magill, M. and Quinzii, M. 2006. *Theory of incomplete markets, Volume I*. Cambridge, MA: MIT Press.

Molico, M. 2006. The distribution of money and prices in search equilibrium. *International Economic Review* 47: 701–722.

Monroe, A. 1966. *Monetary theory before Adam Smith*. New York: Kelley.

Ostroy, J. 1973. The informational efficiency of monetary exchange. *American Economic Review* 63: 597–610.

Patinkin, D. 1951. The invalidity of classical monetary theory. *Econometrica* 19: 134–151.

Redish, A. 2000. *Bimetallism: An economic and historical analysis*. Cambridge: Cambridge University Press.

Renero, J. 1999. Does and should a commodity medium of exchange have relatively low storage costs? *International Economic Review* 40: 251–264.

Samuelson, P. 1961. *Foundations of economic analysis*. Cambridge: Harvard University Press.
Samuelson, P. 1968. What classical and neoclassical monetary theory really was. *Canadian Journal of Economics* 1: 1–15.
Sargent, T., and F. Velde. 2002. *The big problem of small change*. Princeton: Princeton University Press.
Sargent, T., and N. Wallace. 1983. A model of commodity money. *Journal of Monetary Economics* 12: 163–187.
Shi, S. 1995. Money and prices: A model of search and bargaining. *Journal of Economic Theory* 67: 467–498.
Shi, S. 1997. A divisible search model of money. *Econometrica* 65: 75–102.
Shubik, M. 1973. Commodity money, oligopoly, credit and bankruptcy in a general equilibrium model. *Western Economic Journal* 11: 24–38.
Smith, B. 2002. Taking intermediation seriously. *Journal of Money, Credit and Banking* 35: 1319–1358.
Starr, R., and M. Stinchcombe. 1999. Exchange in a network of trading posts. In *Markets, information and uncertainty: Essays in economic theory in honor of Kenneth J. Arrow*, ed. G.Chichilnisky. Cambridge: Cambridge University Press.
Townsend, R. 1989. Currency and credit in a private information economy. *Journal of Political Economy* 97: 1323–1344.
Trejos, A., and R. Wright. 1995. Search, bargaining, money and prices. *Journal of Political Economy* 103: 118–141.
Wallace, N., and T. Zhu. 2007. Float on a note. *Journal of Monetary Economics* 54: 229–246.
Williamson, S., and R. Wright. 1994. Barter and monetary exchange under private information. *American Economic Review* 84: 104–123.
Zhu, T. 2003. Existence of a monetary steady state in a matching model: Indivisible money. *Journal of Economic Theory* 112: 307–324.
Zhu, T. 2005. Existence of a monetary steady state in a matching model: Divisible money. *Journal of Economic Theory* 123: 135–160.

Fichte, Johann Gottlieb (1762–1814)

James Bonar

Fichte, though of the first importance as a philosopher, cannot be called an economist. Yet through his philosophy he has indirectly exercised great influence on economists, his system giving in outline the theory of development worked out by Hegel, and applied by certain of Hegel's followers to economic history and theory. Yet the direct influence of Fichte, through his writings on social and political questions, has been much less strong than might have been expected from the power of the writer and the brilliancy of his theories.

Fichte himself had two social ideals. (*a*) He looked forward to a condition of human society when the state and the coercion of laws would not be needed; as regards the remote future, he is what is now called an anarchist, of the type of William Godwin. (*b*) But he sees that men have, strictly speaking, no rights without the state, and conceives that they must necessarily pass through a stage of development in which the state and the laws shall educate them. He has, therefore, a proximate ideal, an ideal state. The best state is to him a 'closed state'; it is not merely to have its separate nationality and laws, but it is to be separate in its industry and wealth. It is not to be merely 'protected' against its neighbours' competition; it is to have a cordon drawn round it, and, with a few jealously-watched exceptions, it is to have no trade and hardly any intercourse with the foreigner.

The cordon once drawn, the guardians of the state can, he thinks, regulate production and trading, prices and wages. They can introduce a *Landesgeld* or peculiar national currency, valueless abroad; and they can control its value by controlling its quantity. Thus in all departments of economical life there would be hope of introducing constancy, security, and the maintenance of the chief right of man, the right to labour. Fichte means by right to labour the same sort of exclusive privilege as was secured by the old gilds to their members; and he regards this as the most important form of property. Private property in the ordinary sense of the world, family life, and even accumulation of fortunes, are not excluded; and the advantages of family life are clearly recognized. Fichte is a socialist but not a communist; and he does not try to regulate consumption.

The fire of enthusiasm always present in Fichte's writings is not wanting in the *Closed State*; but the *Characteristics*, and *Vocation of Man*, are better examples of his best manner.

His collected works were edited by J.H. Fichte, Berlin, 1845–6 (8 vols). There are passages of

economic interest scattered up and down in nearly all these volumes. *Der Geschlossene Handels-Staat* (1800) was an appendix to the *Naturrecht* (1796). Both are contained in vol. III. of works.

The *Characteristics of the Present Age*, *The Vocation of Man*, and other of the more popular works of Fichte were translated into English (with much spirit) by the late Sir William Smith (Chapman 1848, etc.). The translator published also a *Memoir* of Fichte that went through two editions. Fichte's chief philosophical treatise is *Wissenschaftslehre* (1794), vol. i. of works.

Selected Works

Fichte, Johann Gottlieb. 1845–6. *Collected works*. 8 vols, ed. J.H. Fichte. Berlin.
Fichte, Johann Gottlieb. 1847. *The characteristics of the present age* (trans: Smith, W.). London: Chapman.
Fichte, Johann Gottlieb. 1848. *The vocation of man etc* (trans: Smith, W.). London: Chapman.

Bibliography

Bonar, J. 1893. *Philosophy and political economy*, vol. 4. London.
Lassalle, F. 1862. *Die Philosophie Fichtes und die Bedeutung des deutschen Volksgeistes Festrede*.
Meyer, J.B. 1878. *Fichte, Lassalle, und der Sozialismus*.
Schmoller, G. 1888. *Litteraturgeschichte der Staats- und Socialwissenschaften*. Leipzig.
Smith, W. 1848. *Memoir of Fichte*, 2nd edn. London.

Fictitious Capital

S. De Brunhoff

The concept of 'fictitious capital' is rarely used by economists today. According to the rather small, though diverse, group of authors who have used the notion, it refers to the finance of productive activity by means of credit. Whatever their differences, all authors contrast 'fictitious capital' with 'real capital', where the latter usually refers to produced means of production, but may also include what Marxists call 'money-capital'. One group of authors contrasts finance by means of fictitious capital with voluntary (i.e. not forced) saving of the means of production. Hayek (1939) is a member of this group and refers to Viner's (1937) brief discussion of the use of the concept by English economists (e.g. by Lauderdale and Ricardo). On the other hand, Marx (1894), and Hilferding (1910), analyse the concept of 'fictitious capital' with respect to different forms of 'borrowed capital' and to the significance of the market value of financial titles and their relation to the value produced by labour.

Hayek (1939) argues that fictitious capital is the product of an increase in bank credit which distorts the capital market. When the plans of consumers and entrepreneurs coincide, the credit offered by the former to the latter corresponds to the placement of savings, and the stability of the capital market is assured. However, an increase in bank credit which encourages entrepreneurs to invest without a corresponding increase in saving results in what Hayek calls a crisis of 'over consumption', with, at the same time, a scarcity of capital and an excess supply of unused capital goods. Here the notion of 'fictitious capital' has a pejorative character as if it referred to counterfeit money or a *traite de cavalerie*. It is no longer solely the source of an illusory stimulus but a source of distortion and crisis.

Fictitious capital violates the necessary neutrality of money by establishing a direct relationship between banks and enterprises, in place of the banks' intermediary role. The interpretation of this relationship as illusory or harmful is related to a quantitative conception of the supply of money.

Marx (1894) discusses his quite different notion of 'fictitious capital' in the context of his theory of money and credit. According to him, productive capital, the value of which is created by labour, appears in diverse forms – first, that of money-capital, which is necessary for the payment of wages and the purchase of capital-goods. This money-capital, which is owned by a capitalist, may be loaned by a financier to an entrepreneur. Interest is payable, but this is solely a financial revenue derived from gross profit and has no 'natural' character. According to his A–A′

formula (expressing the cycle of loaned capital), 'capital seems to produce money like a pear-tree produces pears', divorced from the process of production and the exploitation of labour. This is why, according to Marx, interest-bearing capital is the most fetishized form of capital.

The notion of 'fictitious capital' derives from that of loaned money-capital. It suggests a principle of evaluation which is opposed to that which is based on labour-value: 'The formation of fictitious capital is called capitalization. Capitalization takes place by calculating the sum of capital which, at the average rate of interest, would regularly yield given receipts of all kinds.' According to Marx, financial revenues regulate the evaluation of all other receipts. It is 'totally absurd' to capitalize wages as if they were a return to 'human capital', and an 'illusion' to do the same with interest on the public debt to which there corresponds no productive investment.

Nevertheless, the issue of bonds provides the right to a part of the surplus which will be created by future work. Hilferding remains faithful to Marx when he states that 'on the stock exchange, capitalist property appears in its pure form... outside the process of production'. Although doubly fetished, in the circuit A–A′ and on the financial markets, this fictitious capital has some real roots – the necessity of there being money-capital, credit and the means of financial circulation as an expression of the functioning of the capitalist mode of production.

Used in these different ways the notion of 'fictitious capital' has often, for various reasons, a pejorative character. Although little used, it is at the centre of major economic problems: the relation between circulation and production, banks and enterprises and, fundamentally, the distribution of income.

References

Hayek, F.A. 1939. Price expectations, monetary disturbances and malinvestments. In *Profits, interest and investment*, ed. F.A. Hayek. London: Routledge.

Hilferding, R. 1910. *Finance capital*. London: Routledge & Kegan Paul, 1981, Pt 2.

Marx, K. 1894. *Capital*, vol. 3, Part V. Moscow: International Publishers, 1967.

Viner, J. 1937. *Studies in the theory of international trade*. London: Harper.

Fiducial Inference

D. A. S. Fraser

Abstract

Fiducial inference introduced the pivotal inversion that is central to modern confidence theory. Initially this provided confidence bounds but later was generalized to give confidence distributions on the paramseter space. For this it came in direct conflict with the then prominent Bayesian approach called inverse probability. Confidence distributions are now however widespread in modern likelihood theory. Recent results from this theory indicate that the developed fiducial confidence approach is giving a consistent statement of where the parameter is with respect to the data, and indeed is consistent with recent Bayesian approaches that allow data dependent priors.

Keywords

Bayesian inference; Confidence theory; Fiducial inference; Frequentist school; Inverse probability; Likelihood; Markov chain Monte Carlo methods

JEL Classifications

E40

In a seminal paper, R.A. Fisher (1930) introduced the notion of fiducial inference as an alternative to what was then called inverse probability. The key step in fiducial inference is pivotal inversion, which is now standard in all of confidence theory. Fisher's example involved four pairs of observations with a concern for the correlation coefficient ρ between observations in a pair. He had available the distribution function $F(r; \rho)$ for the sample

correlation coefficient r, which depends only on the population correlation ρ; and he had an observed correlation value $r^0 = .99$. He did numerical calculations with the distribution function $F(r; \rho)$, which he had himself previously derived. And he then reported (.765, 1) as a 95 per cent interval for ρ. This is fully in accord with current confidence interval theory. In present notation we would write

$$P(r < .99; \rho) = .95 = P(\hat{\rho}_L < \rho; \rho) = P\{\rho \text{ in } (\hat{\rho}_L, 1); \rho\},$$

where the solution of $F(r; \rho)$=.95 for ρ to obtain the parameter lower bound $\hat{\rho}_L = \hat{\rho}_L(r)$ is standard confidence or pivotal inversion applied to the pivot $u = F(r; \rho)$, which of course has a Uniform (0,1) distribution.

But Fisher (for example, 1930; 1933; 1935; 1956) went further and presented a distribution, called a fiducial distribution, for the parameter ρ, which as a density can be used for calculations such as

$$\int_{.765}^{1} f_{\text{fid}}(\rho; r^0) d\rho = .95,$$

and where for the example the density has the form

$$f_{\text{fid}}(\rho; r) = -(\partial/\partial\rho)F(r; \rho);$$

this density agrees with what in recent likelihood theory would be called a confidence distribution.

But Fisher went still further and spoke of fiducial probability rather than just statements for an interval such as confidence level that we would commonly use. This attribution of probability that a parameter lies in the interval (.765, 1) attracted attack from both the inverse probability community at the time and from the more conventional community that would now be called the frequentist, and includes those having philosophical persuasions. As a consequence, many have viewed fiducial probability as wrong, and strong stigmata have been attached to it. This is rather extraordinary, given that the papers by Fisher are seminal for all of confidence theory and differ only in small deviations of presentation and development.

The key aspects of fiducial that evoked criticism are (*a*) that different pivots can lead to different distributions and thus different intervals, (*b*) that marginalization of a parameter distribution to a component parameter can give a distribution that depends on data in a way different from the obvious that would come from that data, and (*c*) that constraints on the parameter can give a distribution without total probability being equal to 1.

The alternative culture when Fisher (1930) introduced fiducial inference was inverse probability (Bayes 1763). For this, the probability at a data point y^0, given as $f(y^0; \theta)$ and now called likelihood (Fisher 1922) and written $L(\theta; y^0)$, is adjusted by a weight function $w(\theta)$ to give the composite

$$w(\theta)L(\theta; y^0)$$

which is then treated as an unnormed density for the parameter. The weight function $w(\theta)$ is chosen based on properties of the model and called by various names, with default prior being the most unassuming. The present rather large community using this approach is a subgroup of the Bayesian community and the approach has come to be called default Bayesian inference rather than inverse probability analysis; it can also be viewed as a routine frequentist use of the frequentist likelihood function coupled with an ad hoc weight function.

This commonly called default Bayesian approach offers great freedom for the development of statistical techniques: take an observed likelihood $L(\theta; y^0)$ based on Fisher's (1922) proposal; attach a convenient weight function $w(\theta)$ to it; and use the composite for inference for θ. With available high-powered computers and Markov Chain Monte Carlo this leads to a wealth of possible analyses, in contrast to rather limited results from earlier frequentist approaches.

But this leads to perhaps the most influential criticism of the fiducial method (Lindley 1958): (*d*) that a fiducial distribution is typically not an inverse probability or default Bayesian posterior.

Curiously, one finds that the default Bayesian approach is subject to precisely the same criticisms (*a*), (*b*), (*c*) that have been attached to the fiducial approach (for example, concerning (*b*),

see Dawid et al. 1973; see also Fraser 1961, 1995). So the fact (d) that a fiducial analysis is not in general a default Bayesian analysis seems a rather hollow criticism by Lindley (1958). And of course default Bayes typically does not lead to intervals that have the confidence property. Moreover, a recently dominant interest within the current Bayesian community (Fraser and Reid 2002) is to have methods that do reproduce in repeated sampling as do confidence intervals. Perhaps the default Bayesian community is rushing in where the frequentist community neglected its own likelihood function.

But perhaps Fisher and his fiducial approach should be given credit for the fundamental contribution of the pivotal inversion, and of giving rise to the universal confidence procedures. The change of name from fiducial to confidence and then the derogation of fiducial seem a rather heavy historical penalty to Fisher and his profound and seminal developments in statistics. Perhaps 'fiducial' did move too quickly, certainly for the times, and did neglect to develop some fine details. But the results are profound; and the default Bayesian community is finding that it cannot ignore in substance the fiducial criticisms (*a*), (*b*), (*c*); and can't avoid the repeated sampling reproducibility that is the foundation of confidence theory (*d*).

But then, how does fiducial inference work in more general contexts, particularly in the light of recent likelihood theory? For each independent coordinate, say, y_i, a pivot $z_i = h_i(y_i; \theta)$ is needed that describes with full deference to continuity how the coordinate y_i measures or provides information on the parameter θ; this pivot needs to be of the same dimension as the variable y_i and of course as implied by its name has a fixed distribution free of θ. If a coordinate is scalar, the pivot is necessarily equivalent to the distribution function $F_i(y_i; \theta)$ for that coordinate; if it is vector then the choice of pivot represents an explicit statement of how that coordinate variable affects the parameter and is taken as a given for the inference process.

Likelihood theory then shows that the full pivot can be re-expressed to third-order accuracy in the moderate deviations region by an equivalent pivot in which the parameter θ of, say, dimension p appears in only p coordinates of the new pivot. The conditional distribution of these p coordinates given the remaining pivot coordinates (which are of course directly observable) gives effectively a new pivot with of course the same dimension as the parameter. This allows for the standard confidence pivotal inversion to produce confidence regions.

If inference focuses on a particular parameter component $\psi(\theta)$ of interest with dimension d, then the recent likelihood theory shows that the interest parameter can be isolated to third order in a d dimensional component of an equivalent pivot, and the marginal model for that pivot is otherwise free of the full parameter and provides third-order confidence regions for the interest parameter. For some background see Fraser and Reid (2001), Fraser et al. (1999), and Fraser (2004).

See Also

- ▶ Empirical Likelihood
- ▶ Fisher, Ronald Aylmer (1890–1962)
- ▶ Maximum Likelihood

Bibliography

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chance. *Philosophical Transactions of the Royal Society of London* 53: 370–418 .Reprinted in Biometrika 45 (1958), 293–315

Dawid, A., M. Stone, and J. Zidek. 1973. Marginalization paradoxes in Bayesian and structural inference. *Journal of the Royal Statistical Society B* 35: 189–233.

Fisher, R. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A* 222: 309–368.

Fisher, R. 1930. Inverse probability. *Proceedings of the Cambridge Philosophical Society* 26: 528–535.

Fisher, R. 1933. The concept of inverse probability and fiducial probability referring to unknown parameters. *Proceedings of the Royal Statistical Society A* 139: 343–348.

Fisher, R. 1935. The fiducial argument in statistical inference. *Annals of Eugenics* 6: 391–398.

Fisher, R. 1956. *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.

Fraser, D. 1961. The fiducial method and invariance. *Biometrika* 48: 261–280.

Fraser, D. 1995. Some remarks on pivotal models and the fiducial arguments in relation to structural models. *International Statistical Review* 64: 231–236.

Fraser, D. 2004. Ancillaries and conditional inference, with discussion. *Statistical Science* 19: 333–369.
Fraser, D., and N. Reid. 2001. Ancillary information for statistical inference. In *Empirical Bayes and likelihood inference*, ed. S. Ahmed and N. Reid. New York: Springer-Verlag.
Fraser, D., and N. Reid. 2002. Strong matching of frequentist and Bayesian inference. *Journal of Statistical Planning and Inference* 103: 263–285.
Fraser, D., N. Reid, and J. Wu. 1999. A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* 86: 249–264.
Lindley, D. 1958. Fiducial distribution and Bayes' theorem. *Journal of the Royal Statistical Society B* 20: 102–107.
Neyman, J. 1941. Fiducial argument and the theory of confidence intervals. *Biometrika* 32: 128–150.

Fiduciary Issue

J. K. Horsefield

The fiduciary issue of a bank (*fiduciarius* = held in trust) is that part of its note issue that is not covered by gold or by some other generally accepted means of payment, such as silver. The expression is associated especially with the Bank of England, where it dates from the Bank Charter Act 1844 (7 & 8 Vict., c.32), although this does not use the actual expression.

Statutory Provisions

The 1844 Act, as a condition of renewing the Bank of England's charter, required it to divide its activities between an Issue Department and a Banking Department. The Issue Department, as its name implies, was responsible for the control of the Bank's note issue, and in particular for ensuring that the size of the issue complied with the Act. This prescribed that, except for a fixed amount of Government securities, the Bank's notes must be covered completely by gold coin, or by gold or silver bullion of which at least four-fifths must be gold. It is the amount of securities so fixed that is known as the fiduciary issue.

In 1844 the amount of the fiduciary issue was set at £14 million. No official reason was given for choosing this amount, but contemporaries offered a number of possible explanations. Firstly, it was probably no coincidence that the Bank's capital was, and is, £14,553,000. Secondly, an internal Bank committee, which reported while the Act was in preparation, suggested that it would be appropriate to issue £12 million of notes plus £2 million against 'unemployable deposits'. Thirdly, some commentators related the figure to the minimum actual circulation of notes, which between 1799 and 1844 had never fallen below about £15.5 million. Alternatively, it was argued that between 1826 and 1843 the average circulation of notes in excess of the Bank's holdings of bullion had been slightly above £1 million. Adding to this £3 million to replace the notes of certain country banks which had ceased to issue notes, produced a figure of £14 million. This was taken to be the amount that, characteristically, the Bank could float and the public could use. It seems probable that the decision to fix the fiduciary issue at £14 million reflected more than one of these converging considerations.

The fiduciary issue is represented by the first two items among the assets.

The Bank was required by the Act to publish weekly a Return showing how it was complying with the obligations placed upon it. In the first such Return published, for the week ended 7 September 1844, the part concerned with the Issue Department was as shown in Table 1.

The Act also restricted the issue of notes by banks other than the Bank of England. Only those banks already issuing notes on 6 May 1844 might do so in England in future, and the amount which each might issue was limited to those in circulation on that date. Two other provisions in the Act, continuing restrictions already in force, ensured that in the course of time all English note issues other than those of the Bank of England would disappear. These provisions were that no issuing bank might have more than six partners, and that no bank in London or within 65 miles of London (except of course the Bank of England itself) might issue notes.

Fiduciary Issue, Table 1

Issue department			
	£		£
Notes issued	28,351,295	Government debt	11,015,100
		Other securities	2,984,900
		Gold coin and bullion	12,657,208
		Silver bullion	1,694,087
	28,351,295		28,351,295

Increases in Limit

When, as a result of such restrictions, a country bank ceased to issue notes, the Bank of England was permitted to seek authority to increase its fiduciary issue by an amount equal to two-thirds of that which had lapsed. (The limitation to two-thirds appears to have been based on an assumption that the discontinuing bank would normally have held a reserve in gold or Bank of England notes equal to one-third of its note issue.) As a result of this provision, the Bank of England's issue not covered by coin or bullion increased by stages, eventually reaching £19,750,000 on 21 February 1923.

During World War I the issue of bank notes was supplemented by Government-issued Treasury Notes, but in 1928 the two series were amalgamated under the aegis of the Bank of England. The operative statute was the Currency and Bank Notes Act 1928 (18 & 19 Geo.V, c.13). This set the limit of the fiduciary issue at £260 million, but included provision for this to be increased or decreased on the initiative of the Bank of England. Increases, which might continue for six months at a time, were to be authorized by Treasury Minute, and were subject to a maximum of two years, after which parliamentary approval had to be obtained. Reductions, on the other hand, could be authorized by a Treasury letter. The backing for the fiduciary issue was still to be Government debt, except that silver coin to an amount not exceeding £5,500,000 might be included. Apart from the fiduciary issue, all notes had to be covered by gold coin or bullion.

The crisis of 1931, leading to Great Britain's abandonment of the Gold Standard, was accompanied by an increase in the fiduciary issue to £275 million, which was in force from August 1931 to March 1933. Thereafter the limit varied between £200 million and £260 million until 1939. In January 1939 it was temporarily increased to £400 million by a Treasury Minute. In March of that year it was altered to £300 million by the Currency and Bank Notes Act 1939 (2 & 3 Geo.VI, c.7). In September 1939, however, practically the whole of the Bank's gold holding was transferred to the Exchange Equalization Account, and the fiduciary issue was increased to £580 million. Since then the Bank's note issue has been effectively backed only by paper. At the end of the Bank's year 1983–4 (February 1984) the notes issued totalled £11,470,000,000, while the assets of the Issue Department consisted wholly of securities. The increase above the £300 million set by the Currency and Bank Notes Act 1939 is authorized regularly by the Treasury, and is confirmed by Statutory Order placed before Parliament every second year.

Purpose of Limitation

The philosophy underlying the limitation of the fiduciary issue was that of the Currency School. It was held that to restrict the issue of bank notes in this way would ensure that there would be no repetition of the crisis of 1836, which was believed to have been caused by an undue proliferation of notes. For this reason, proposals put forward while the Act was being deliberated, by which a relaxing clause would have been included to allow for emergencies, were held to be unnecessary, and indeed unwise. Subsequent experience, however, ensured that such a clause was included when the 1928 Act was being drafted. For, far from preventing new crises, the 1844 Act in some respects promoted them by leading the Bank to believe that it was fulfilling its responsibilities if the note issue was within the prescribed limit, without regard to the ability of the Banking Department to expand credit.

The outcome was a series of crises in 1847, 1857 and 1866. On each occasion commercial panics produced scrambles for liquidity, which led inevitably to demands for more Bank of England notes. Each time, the Bank was initially prevented from responding by the limit on its note issue, thereby exaggerating the panic. Each time, however, the Government encouraged the Bank to meet commercial requirements, even though the volume of notes issued might exceed the statutory limit and undertook to indemnify the Bank if this occurred. In practice, the limit was not exceeded in 1847 or 1866, but in 1857 the note issue was increased by £2 million above the £14,475,000 which was then the fiduciary issue; of these £2 million, some £928,000 left the Bank.

These developments, revealing the inadequacy of the Currency Theory, cast doubts on the significance of the note issue, and therefore of the limit to the fiduciary issue. Concern also shifted to the size of the Bank's gold stock in relation to the country's international commitments.

The Macmillan Committee, reporting in 1931 (paragraph 328), recommended that the fiduciary issue as such should be abolished, being replaced by a limit on the Bank's total note issue, together with an obligation to maintain a minimum stock of gold. This proposal was not adopted.

In 1959 the Radcliffe Committee, whose report stressed that it was the money supply as a whole rather than the note issue which was important, dismissed the fiduciary issue as irrelevant. The Committee further remarked (paragraph 367) that the only current use of the Bank Return, as prescribed in 1844, was to 'provide a formula for determination of the income of which the Bank has untrammelled disposal'.

Today the fiduciary issue would appear to have no other function than, through the two-year limitation upon its increase imposed in 1928, to afford Parliament a periodic reminder of the growth of the monetary base.

Scotland and Ireland

In Scotland and Ireland the individual banks have continued to issue notes, there being no equivalent there to the provision in the Bank Charter Act 1844 extinguishing English note issues other than those of the Bank of England. However, in 1845 limits similar in effect to the Bank of England's fiduciary issue were placed upon the volume of notes which each of the 19 Scottish banks might issue, other than against a backing of legal tender (8 & 9 Vict., c.38). These limits, which totalled some £3 million for Scotland as a whole, were based on the average of each bank's actual circulation during the twelve months ended 1 May 1845. Similar legislation was passed for Ireland (8 & 9 Vict., c.37).

In 1928 parallel legislation to the Currency and Bank Notes Act restricted the fiduciary issues of the Scottish banks (then numbering eight) to a total of £2,676,350 and those of the banks in Northern Ireland to £1,634,000.

See Also

- ▶ Banking School, Currency School, Free Banking School
- ▶ Monetary Base

Bibliography

Clapham, Sir John. 1944. *The bank of England: A history.* Cambridge: Cambridge University Press.

Committee on Finance and Industry (Macmillan Committee). 1931. *Report* Cmd 3897. London: HMSO.

Committee on the Working of the Monetary System (Radcliffe Committee). 1959. *Report* Cmnd 827. London: HMSO.

Sayers, R.S. 1976. *The Bank of England, 1891–1944*. Cambridge: Cambridge University Press.

Field Experiments

John A. List and David Reiley

Abstract

Field experiments have grown significantly in prominence since the 1990s. In this article, we

provide a summary of the major types of field experiments, explore their uses, and describe a few examples. We show how field experiments can be used for both positive and normative purposes within economics. We also discuss more generally why data collection is useful in science, and more narrowly discuss the question of generalizability. In this regard, we envision field experiments playing a classic role in helping investigators learn about the behavioural principles that are shared across different domains.

Keywords

Charitable giving; Field experiments; Generalizability; Laboratory experiments; Matching funds; Testing; Uniform-price auctions; Vickrey auctions

JEL Classifications

C1

Field experiments occupy an important middle ground between laboratory experiments and naturally occurring field data. The underlying idea behind most field experiments is to make use of randomization in an environment that captures important characteristics of the real world. Distinct from traditional empirical economics, field experiments provide an advantage by permitting the researcher to create exogenous variation in the variables of interest, allowing us to establish causality rather than mere correlation. In relation to a laboratory experiment, a field experiment potentially gives up some of the control that a laboratory experimenter may have over her environment in exchange for increased realism.

The distinction between the laboratory and the field is much more important in the social sciences and the life sciences than it is in the physical sciences. In physics, for example, it appears that every hydrogen atom behaves exactly alike. Thus, when astronomers find hydrogen's signature wavelengths of light coming from the Andromeda Galaxy, they use this information to infer the quantity of hydrogen present there. By contrast, living creatures are much more complex than atoms and molecules, and they correspondingly behave much more heterogeneously. Despite the use of 'representative consumer' models, we know that not all consumers purchase the same bundle of goods when they face the same prices. With complex, heterogeneous behaviour, it is important to sample populations drawn from many different domains – both in the laboratory and in the field. This permits stronger inference, and one can also provide an important test of generalizability, testing whether laboratory results continue to hold in the chosen field environment.

We find an apt analogy in the study of pharmaceuticals, where randomized experiments scientifically evaluate new drugs to treat human diseases. Laboratory experiments evaluate whether drugs have desirable biochemical effects on tissues and proteins *in vitro*. If a drug appears promising, it is next tested *in vivo* on several species of animals, to see whether it is absorbed by the relevant tissues, whether it produces the desired effects on the body, and whether it produces undesirable side effects. If it remains with significant promise after those tests, it is then tested in human clinical trials to explore efficacy and measure any side effects.

Even after being tested thoroughly in human clinical trials and approved by regulators, a drug may sometimes reveal new information in large-scale use. For example, *effectiveness* may be different from the *efficacy* measured in clinical trials: if a drug must be taken frequently, for example, patients may not remember to take it as often as they are supposed to or as often as they did in closely supervised clinical trials. Furthermore, rare side effects may show up when the drug is finally exposed to a large population.

Much like this stylized example, in economics there are a number of reasons why insights gained in one environment might not perfectly map to another. Field experiments can lend insights into this question (see also Bohm 1972; Harrison and List 2004; Levitt and List 2006; List, 2007). First, different types of subjects might behave differently; university students in the laboratory might not exhibit the same behaviour as financial traders or shopkeepers. In particular, the people who undertake a given economic activity have selected into that activity and market forces might have

changed the composition of players as well; you might expect regular bidders to have more skill and interest in auctions than a randomly selected laboratory subject, for example.

A second reason why a field experiment might differ from a laboratory experiment is that the laboratory environment might not be fully representative of the field environment. For example, a typical donor asked to give money to charity might behave quite differently if asked to participate by choosing how much money to contribute to the public fund in a public-goods game (List, 2007). The charitable- giving context could provide familiar cognitive cues that make the task easier than an unfamiliar laboratory task. Even the mere fact of knowing that one's behaviour is being monitored, recorded, and subsequently scrutinized might alter choices (Orne 1962).

Perhaps most important is the fact that any theory is an approximation of reality. In the laboratory, experimenters usually impose all the structural modelling assumptions of a theory (induced preferences, trading institutions, order of moves in a game) and examine whether subjects behave as predicted by the model. In a field experiment, one accepts the actual preferences and institutions used in the real world, jointly testing both the structural assumptions (such as the nature of values for a good) and the behavioural assumptions (such as Nash equilibrium).

For example, Vickrey (1961) assumes that in an auction there is a fixed, known number of bidders who have valuations for the good drawn independently from the same (known) probability distribution. He uses these assumptions, along with the assumption of a risk-neutral Nash equilibrium, to derive the 'revenue equivalence' result: that Dutch, English, first-price, and second-price auctions all yield the same expected revenue. However, in the real world the number of bidders might actually vary with the good or the auction rules, and the bidders might not know the probability distribution of values. These exceptions do not mean that the model should be abandoned as 'wrong'; it might well still have predictive power if it is a reasonable approximation to the truth. In a field experiment (such as Lucking-Reiley 1999, for this example), we approach the real world; we do not take the structural assumptions of a theory for granted.

Such an example raises the natural question related to the actual difference between laboratory and field experiments. Harrison and List (2004) propose six factors that can be used to determine the field context of an experiment: the nature of the subject pool, the nature of the information that the subjects bring to the task, the nature of the commodity, the nature of the task or trading rules applied, the nature of the stakes, and the environment in which the subjects operate. Using these factors, they discuss a broad classification scheme that helps to organize one's thoughts about the factors that might be important when moving from the laboratory to the field.

A first useful departure from laboratory experiments using student subjects is simply to use 'non-standard' subjects, or experimental participants from the market of interest. Harrison and List (2004) adopt the term 'artefactual' field experiment to denote such studies. While one might argue that such studies are not 'field' in any way, for consistency of discussion we denote such experiments as artefactual field experiments for the remainder of this article, since they do depart in a potentially important manner from typical laboratory studies. This type of controlled experiment represents a useful type of exploration beyond traditional laboratory studies.

Moving closer to how naturally occurring data are generated, Harrison and List (2004) denote a 'framed field experiment' as the same as an artefactual field experiment but with field context in the commodity, task, stakes, or information set of the subjects. This type of experiment is important in the sense that a myriad of factors might influence behaviour, and by progressing slowly towards the environment of ultimate interest one can learn about whether, and to what extent, such factors influence behaviour in a case-by-case basis.

Finally, a 'natural field experiment' is the same as a framed field experiment but where the environment is one where the subjects naturally undertake these tasks and where the subjects do not know that they are participants in an experiment. Such an exercise represents an approach that combines the most attractive elements of the laboratory

and naturally occurring data – randomization and realism. In this sense, comparing behaviour across natural and framed field experiments permits crisp insights into whether the experimental proclamation, in and of itself, influences behaviour.

Several examples of each of these types of field experiments are included in List (2006). Importantly for our purposes, each of these field experimental types represents a distinct manner in which to generate data. As List (2006) illustrates, these field experiment types fill an important hole between laboratory experiments and empirical exercises that make use of naturally occurring data. Yet an infrequently discussed question is: why do we bother to collect data in economics, or in any science?

First, we use data to collect enough facts to help construct a theory. Several prominent broader examples illustrate this point. After observing the anatomical and behavioural similarities of reptiles, one may theorize that reptiles are more closely related to each other than they are to mammals on the evolutionary tree. Watson and Crick used data from Rosalind Franklin's X-ray diffraction experiment to construct a theory of the chemical structure of DNA. Careful observations of the motions of the planets in the sky led Kepler to theorize that planets (including Earth) all travel in elliptical orbits around the Sun, and Newton to theorize the inverse-square law of gravitation. After observing with a powerful telescope that the fuzzy patches called 'spiral nebulae' are really made up of many stars, one may theorize that our solar system is itself part of its own galaxy, and the spiral nebulae are external to our Milky Way galaxy. Robert Boyle experimented with different pressures using his vacuum pump in order to infer the inverse relationship between the pressure and the volume of a gas. Rutherford's experiments of shooting charged particles at a piece of gold foil led him to theorize that atoms have massive, positively charged nuclei.

Second, we use data to test theories' predictions. Galileo experimented with balls rolling down inclined planes in order to test his theory that all objects have the same rate of acceleration due to gravity. Pasteur rejected the theory of spontaneous generation with an experiment that showed that microorganisms grow in boiled nutrient broth when exposed to the air, but not when exposed to carefully filtered air. Arthur Eddington measured the bending of starlight by the sun during an eclipse in order to test Einstein's theory of general relativity.

Third, we use data to make measurements of key parameters. On the assumption that the electron is the smallest unit of electric charge, Robert Millikan experimented with tiny, falling droplets of oil to measure the charge of the electron. On the assumption that radioactive carbon-14 decays at a constant rate, archaeologists have been able to provide dates for various ancient artifacts. Similarly, scientists have assumed theory to be true and designed careful measurements of many other parameters, such as the speed of light, the gravitational constant, and various atomic masses.

Field experiments can be a useful tool for each of these purposes. For example, Anderson and Simester (2003) collect facts useful for constructing a theory about consumer reactions to nine-dollar endings on prices. They explore the effects of different price endings by conducting a natural field experiment with a retail catalogue merchant. Randomly selected customers receive one of three catalogue versions that show different prices for the same product. Systematically changing a product's price varies the presence or absence of a nine-dollar price ending. For example, a cotton dress may be offered to all consumers, but at prices of 34, 39, and 44 dollars, respectively, in each catalogue version. They find a positive effect of a nine-dollar price on quantity demanded, large enough that a price of 39 dollars actually produced higher quantities than a price of 34 dollars. Their results reject the theory that consumers turn a price of 34 dollars into 30 dollars by either truncation or rounding. This finding provides empirical evidence on an interesting topic and demonstrates the need for a better theory of how consumers process price endings.

List and Lucking-Reiley (2000) present an example of a framed field experiment designed to test a theory. The theory of multi-unit auctions predicts that a uniform-price sealed-bid auction will produce bids that are less than fully demand-revealing, because such bids might lower the price paid by the same bidder on another unit.

By contrast, the generalized Vickrey auction predicts that bidders will submit bids equal to their values. In the experiment, List and Lucking-Reiley conduct two-person, two-unit auctions for collectible sportscards at a card trading show. The uniform-price auction awards both items to the winning bidder(s) at an amount equal to the third-highest bid (out of four total bids), while the Vickrey auction awards the items to the winning bidder(s) for amounts equal to the bids that they displaced from winning. List and Lucking-Reiley find that, as predicted by the theory of demand reduction, the second-unit bids submitted by each bidder were lower in the uniform-price treatment than in the Vickrey treatment. The first-unit bids were predicted to be equal across treatments, but in the experiment they find that the first-unit bids were anomalously higher in the uniform-price treatment. Subsequent laboratory experiments (see, for example, Engelmann and Grimm 2003; Porter and Vragov 2003), have confirmed this finding.

Finally, Karlan and List (2007) is an example of a natural field experiment designed to measure key parameters of a theory. In their study, they explore the effects of 'price' changes on charitable giving by soliciting contributions from more than 50,000 supporters of a liberal organization. They randomize subjects into several different groups to explore whether solicitees respond to upfront monies used as matching funds. They find that simply announcing that a match is available considerably increases the revenue per solicitation – by 19 per cent. In addition, the match offer significantly increases the probability that an individual donates – by 22 per cent. Yet, while the match treatments relative to a control group increase the probability of donating, larger match ratios – 3:1 dollars (that is, 3 dollars match for every 1 dollar donated) and 2:1 dollar – relative to smaller match ratios (1:1 dollar) have no additional impact.

In closing, we believe that field experiments will continue to grow in popularity as scholars continue to take advantage of the settings where economic phenomena present themselves. This growth will lead to fruitful avenues, both theoretical and empirical, but it is clear that regardless of the increase in popularity, the various empirical approaches should be thought of as strong complements, and combining insights from each of the methodologies will permit economists to develop a deeper understanding of our science.

See Also

- ▶ Experimental Economics
- ▶ Experimental Economics, History of
- ▶ Experimental Labour Economics
- ▶ Experimental Methods in Economics
- ▶ Experimental Methods in Environmental Economics
- ▶ Experiments and Econometrics

Bibliography

Anderson, E.T., and D. Simester. 2003. Effects of $9 price endings on retail sales: Evidence from field experiments. *Quantitative Marketing and Economics* 1: 93–110.

Bohm, P. 1972. Estimating the demand for public goods: An experiment. *European Economic Review* 3: 111–130.

Engelmann, D. and V. Grimm. 2003. Bidding behavior in multi-unit auctions – An experimental investigation and some theoretical insights. Working paper. Prague: Centre for Economic Research and Graduate Education, Economic Institute.

Harrison, G.W., and J.A. List. 2004. Field experiments. *Journal of Economic Literature* 42: 1009–1055.

Karlan, D. and J.A. List. 2007. Does price matter in charitable giving? Evidence from a large-scale natural field experiment. *American Economic Review,* forthcoming.

Levitt, S.D., and J.A. List. 2006. What do laboratory experiments measuring social preferences tell us about the real world? *Journal of Economic Perspectives* 21(2): 153–174.

List, J.A. 2006. Field experiments: A bridge between lab and naturally occurring data. *Advances in Economic Analysis & Policy* 6(2), Article 8. Abstract online. http://www.bepress.com/beieap/advances/vol6/iss2/art8. Accessed 26 May 2007.

List, J.A., and D. Lucking-Reiley. 2000. Demand reduction in a multi-unit auction: Evidence from a sportscard field experiment. *American Economic Review* 90: 961–972.

Lucking-Reiley, D. 1999. Using field experiments to test equivalence between auction formats: Magic on the Internet. *American Economic Review* 89: 1063–1080.

Orne, M.T. 1962. On the social psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist* 17: 776–783.

Porter, D. and R. Vragov. 2003. An experimental examination of demand reduction in multi-unit versions of the uniform-price, Vickrey, and English auctions. Working

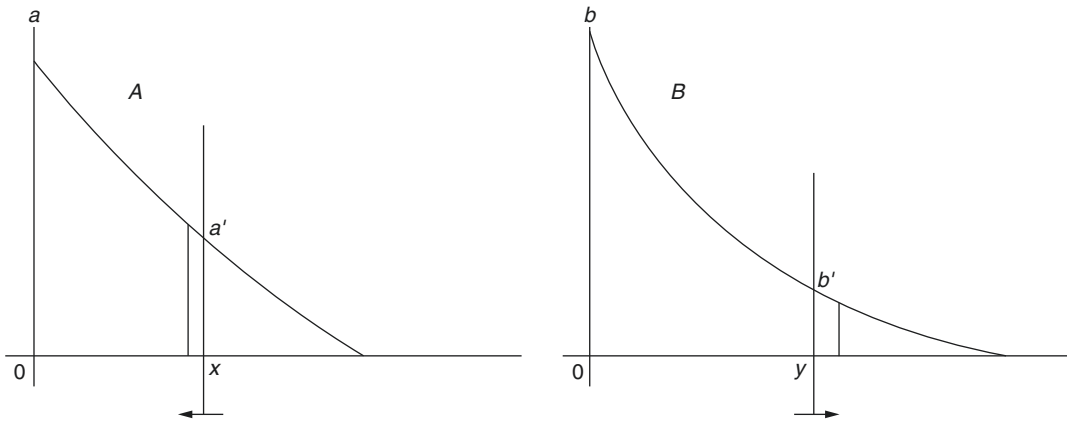
paper. Interdisciplinary Center for Economic Science, George Mason University.
Vickrey, W. 1961. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance* 16: 8–37.

Final Degree of Utility

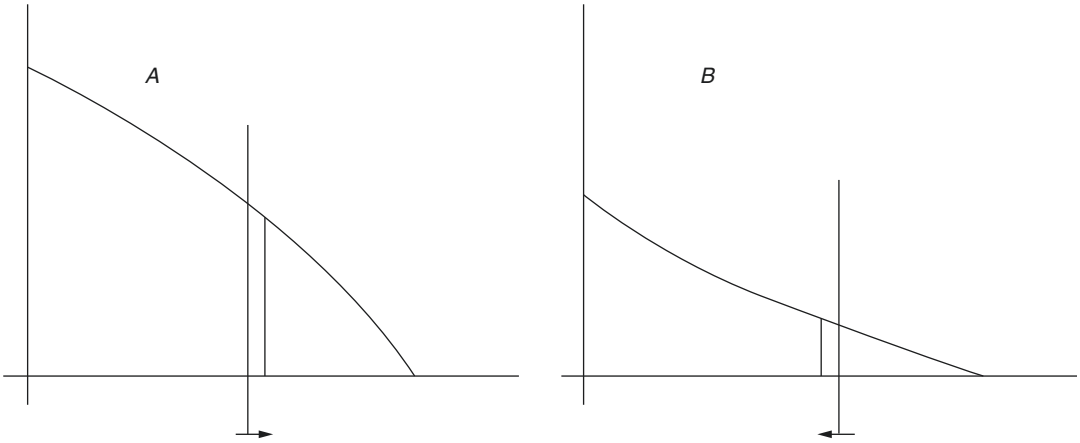
P. H. Wicksteed

The expression used by Jevons for the degree of utility of the last increment of any commodity secured, or the next increment expected or desired. The increments being regarded as infinitesimal, the degree of utility is not supposed to vary from the last possessed to the next expected. It will be obvious, after a study of the article on Degree of Utility that it is the *final* degree of utility of various commodities that interests us commercially, not, for instance, their initial or average degrees of utility. That is to say (Fig. 1), if a is a small unit of the commodity A, and b a small unit of the commodity B, and q_a the quantity of A I possess, and q_b the quantity of B I possess, then, in considering the equivalence of a and b I do not ask whether A or B has the greater initial degree of utility, i.e. I do not compare the lines Oa and Ob, nor do I inquire which has the greater average degree of utility, i.e. I do not compare the height of the rectangle on base Ox which shall equal the area aOxa', with the height of the rectangle on base Oy which shall equal the area bOyb', but I compare the length xa' with the length yb', and ask what are the relative rates at which increments of A and B will *now add* to my satisfaction. If xa' is twice the length of yb', then (since a and b are supposed to be small units, throughout the consumption of which the decline in the curves aa' bb' may be neglected) it is obvious that 2b will be equivalent to a, since either increment will yield an equal area of satisfaction.

Now suppose (Fig. 2) that some other possessor of the commodities A and B, either because he possesses them in different proportions, or because his tastes and wants are different, finds that the relative final utilities of the small units a and b are not the same for him (2) as they are for me (1). Say that for him 3b is the equivalent of a, clearly the conditions for a mutually advantageous exchange exist. Let δ be greater than 2 and less than 3, so that δ - 2 and 3 - δ are both positive. Now suppose (1) exchanges with (2), giving him a and receiving from him δb. Then, (1) receives δb in exchange for a (worth 2b to him) and benefits to the extent of (δ – 2)b, and by the same transaction (2) has received a (worth 3b to him) in exchange for δb, and has benefited to the extent of (3 – δ)b. The result of this exchange will be a movement of all the verticals that indicate the amount of each commodity possessed by each exchanger, in the directions indicated by the arrow-heads; and this again will (as is obvious



Final Degree of Utility, Fig. 1



Final Degree of Utility, Fig. 2

from inspection of the figures) tend to reduce the difference between the ratio of equivalence between *a* and *b* in the case of the two exchangers. The process of exchange will go on (δ not necessarily remaining constant) until the ratio of equivalence between *a* and *b* coincides for the two exchangers, the last exchange bringing about an equilibrium in accordance with that ratio. Such a ratio of equilibrium is a limiting ratio of exchange; that is to say, exchange constantly tends to approach such a ratio, perhaps by a series of tentative exchanges at various rates, and would cease were such a ratio actually arrived at.

Hence Jevons's fundamental theorem: 'The ratio of exchange of any two commodities will be the reciprocal of the ratio of the final degrees of utility of the quantities of commodities available for consumption after the exchange is completed', applies to an ideal ratio which would secure equilibrium at a stroke, rather than to the tentative bargains by which it is approached in the 'actual market'.

The conceptions of 'degree of utility' and 'final degree of utility' lie at the heart of the mathematical method of political economy, and their complete history would almost coincide with the history of mathematical economics. Incidentally the idea has been struck from time to time by sundry mathematicians, and it has been worked out independently by economists no fewer than four or five times. Cournot (1838), Dupuit (1844), Gossen (1853), and Jevons (1862 and 1871) successively discovered and taught the theory, each one in ignorance of the work of his predecessors. In 1871 the Austrian Menger, and in 1874 the Swiss Walras (working on the basis laid down by Cournot), adopted essentially the same central conception, and since then the theory has not again sunk into oblivion. Many writers in Germany, Holland, Denmark, France, Italy, and England are now engaged in developing it. See the bibliographies and lists of writers in the appendix to Jevons's *Theory of Political Economy*, 3rd edn, and the Preface to Walras's *Théorie de la Monnaie*, 1886; and for far-reaching recent developments in America, England, and France see Appendix.

[Jevons's 'final degree of utility' is the *Grenznutzen* of the Austrian school, Gossen's *Werth der letzten Atome*, and Walras's *rareté*.]

Reprinted from *Palgrave's Dictionary of Political Economy*

Bibliography

Gossen, H.H. 1854. *Entwicklung der Gesetze des menschlichen Verkehrs, und der daraus fliessenden Regeln für menschliches Handeln*. Braunschweig: Vieweg.

Jevons, W.S. 1888. *1871. Theory of political economy. London. 3rd edn*. London: Macmillan.

Walras, L. 1886. *Théorie de la monnaie*: Lausanne: Corbaz.

Final Utility

P. H. Wicksteed

The principles and methods embodied in Jevons's doctrine of 'final utility' have received farreaching developments in recent years. Hence a movement has arisen, variously described as 'psychological' or 'marginalist', which aims at unifying and simplifying economic theory, and at the same time affiliating its laws more closely to the principles that regulate human conduct in general.

Jevons has shown that the demand in a market in which there are no reserved prices can be represented by a collective curve. The amount of the commodity in the market is measured on the abscissa, and the equilibrating price on the ordinate. The next step is to point out that in so far as the sellers have reserved prices they ought to be regarded as themselves entering the market, with potential demands, on the same footing as the purchasers. Their intention to retain such and such quantities of their stock at such and such prices (whether for their own use or because they speculate on the demands of future purchasers) constitute *de facto* demands, and should be entered on the collective demand curve; which, together with the register of the amount of the commodity, will determine the price, as before. It follows that the cross curves of demand and supply, so often employed by economists, are really no more than two sections of the true collective curve of demand, separated out from each other, and read, for convenience, in reverse directions. This separation is irrelevant to the determination of the equilibrating price (as may easily be shown by experiment), though it enables us to read off the volume of the exchanges that will be necessary in order to bring about the equilibrium, on any given supposition as to initial holdings. These cross curves, then, as usually presented, confuse the methods by which the equilibrating price is arrived at with the conditions that determine what it is.

Passing on to the problems of production and distribution, we note that in an industrially advanced community production rests upon the cooperation of a number of heterogeneous factors, the supply of which may be controlled by a number of independent individuals or combinations; and since it is obvious that the value of a means of production must be derivative from the value of the product, we have, theoretically, to determine the principle on which the value of the product when realized will be distributed amongst the various factors which cooperated in its production. Practically the factors will generally be brought together by a series of speculative transactions based on estimates made in advance. But in any case the value of the several factors must be determined by consideration of their productive effectiveness at the margin, and their equivalence to each other in fractional substitutions. For although the nature of the productive service rendered by such factors as land, labour, and tools, for instance, is different in each case, and no main factor could be replaced in its entirety by any other, yet every manager is constantly engaged in considering alternatives and equivalences between fractional additions or subtractions of them at the margin. It is so that he determines the proportions in which to distribute his resources over the improving or extending of a site, the modification of existing buildings, the replacing of machinery, the strengthening or reduction of this or that grade of labour, superintendence to reduce the waste of raw material, or the seeking of new openings, or maintenance of old ones, by advertisement. And all the time he has to convince his employers that his own skill in judging of these matters is as effectively productive as any increments in the more immediate factors of production that they could command for the salary that they pay him. The purchasers, then, in the great markets of the productive factors consider them under the uniform aspect of their relative productive efficiency at the margin, just as the purchaser in the retail market considers his heterogeneous purchases under the uniform aspect of their relative efficiency at the margin, in gratifying his desires or expressing his impulses. In a word, there are not many laws of

distribution but one, and that law is the law of the market.

(Thus 'interest' is the price, reckoned in deferred payments, of present command of resources. The industrial, who expects this command actually to produce the future resources out of which he will make the payment, enters a market in which he will have to compete with the non-industrial who is willing to risk or compromise his future at the dictate of his present desires, and the ordinary consumer who, having a small revenue and no accumulations, is willing to pay a higher price for a possession, if he may spread the payment over a longer period, rather than cut deep into the quick of his other requirements at the moment.

('Rent' is a form of hire, the continuous purchase of a continuous revenue of services or enjoyments. The well-known figure of the rent curve, which represents the decreasing productive efficiency of successive applications of labour and capital to a fixed unit of land, is seen to owe its form not to any special characteristic of land but to the selection of a single factor of production which is not to increase while all the others do. The identical facts which such a curve represents, if read in the reverse order, would represent the same series of hypotheses as to the relative proportions of the several factors; but the rent would now be presented as a rectangular area, with its altitude determined by the alternative uses of land, and the return to labour and capital, as a curvilinear 'residue', determined by the decreasing yield of a fixed constant of labour, etc., when spread over more and more land.)

Thus it will be seen that the end dominates the means throughout. The direction and administration of all resources is ultimately determined by estimates of the value of some experience, or by the imperativeness of some expression of the human consciousness. If at any point the expectations based on these estimates should fail or wither, the breadth of the stream that has already flowed at their bidding is powerless to sustain their living significance. Anticipated value determines the cost and sacrifice that will be incurred in production, but the cost and sacrifice, when once incurred, cannot control the value of the product.

If we now return to our starting-point in Jevons's 'final utility' and its control of the distribution of a man's pecuniary resources, we note that the term 'final' has been generally abandoned. It seems to imply a succession of experiences, following each other in time, as when a man's hunger is gradually appeased and each morsel meets a decreasingly urgent need. It is therefore inapplicable, for instance, to the problems we have discussed under the head of 'distribution', where the units of the same factor may be indistinguishable in quality and may all be running abreast of each other in the output of a continuous stream of efficiency, but where nevertheless the withdrawal from cooperation of one unit out of five would be a less serious matter than the withdrawal of one out of four, because it would create a less serious disturbance of the proportions between the factors and would require less serious readjustments or additions to compensate it. The term 'marginal' has been very generally adopted, but it has the disadvantage of still suggesting (especially in connection with land) some intrinsic differentiating characteristic which earmarks and individualizes a unit as 'marginal' in virtue of its own nature. The term 'fractional' may often be conveniently used.

Again, the world 'utility' so conspicuously fails to include all the objects of wise or foolish, good or bad desire, to which the economic machinery ministers, that if it still sometimes retains its place (subject to careful explanation that it does not really *mean* utility) it is only for want of general agreement as to a substitute. The anomaly becomes more glaring and extends to the term 'consumption', when we realize that the laws of political economy are but the application to a special set of problems of the universal laws of the distribution and administration of resources in general (whether of money, time, influence, powers of thought, or aught else) amongst all the objects that we deliberately pursue or to which we are spontaneously impelled, whether material or spiritual, private or social, wise or foolish. It is intolerable that 'consumption' (with its subtle suggestion of a regrettable

necessity that puts a drag upon the progress of 'production') should continue to stand for the whole stream of 'actualizings', in conscious experience, of the potentialities to the development of which human effort is devoted. It is the nature of these actualizings, contemplated or realized, that is the supremely significant thing in the life of a man or a community; for it is from them that all which leads up to them derives its worth or its worthlessness.

Finance

Stephen A. Ross

Abstract

The neoclassical theory of finance is based on the study of (*a*) efficient markets, meaning markets that use all available information in setting prices, (*b*) the trade-off between return and risk, (*c*) option pricing and the principle of no arbitrage, and (*d*) corporate finance, that is, the structure of financial claims issued by companies. This article surveys these theories and their empirical support and it also identifies certain empirical regularities unexplained by the neoclassical theory that are being addressed by theories of asymmetric information.

Keywords

Arbitrage; Arbitrage pricing theory; Arrow-Debreu model; Asymmetric information; Black-Scholes model; Capital asset pricing model; Consumption beta model; Corporate finance; Corporate taxation; Efficiency; Efficient markets hypothesis; Finance; Hicks, J.; Interest rates; Martingales; Mean variance analysis; Merton, R.; Modigliani-Miller theorem; Option pricing; Pareto efficiency; Principal and agent; Productive efficiency; Rational expectations; Risk and return; Risk premium; Security market line equation; Signalling models; Tobin, J.; von Neumann-Morgenstern utility function

JEL Classifications

C9

Finance is a subfield of economics distinguished by both its focus and its methodology. The primary focus of finance is the workings of the capital markets and the supply and the pricing of capital assets. The methodology of finance is the use of close substitutes to price financial contracts and instruments. This methodology is applied to value instruments whose characteristics extend across time and whose payoffs depend upon the resolution of uncertainty.

Finance is not terribly concerned with the problems that arise in a barter economy or, for that matter, in a static and certain world. But, once the element of time is introduced, transactions develop a dual side to them. When a loan is made, the amount and the terms are recorded to insure that repayment can be enforced. The piece of paper or the computer entry that describes and legally binds the borrower to repay the loan can now trade on its own as a 'bearer' instrument. It is at the point when debts were first traded that capital markets and the subject of finance began.

The study of finance is enriched by having a large body of evolving data and market lore and some powerful and, at times, competing intuitions. These intuitions are used to structure our understanding of the data and the markets which generate it. The modern tradition in finance began with the development of well-articulated models and theories to explore these intuitions and render them susceptible to empirical testing.

While the subject of finance is anything but complete, it is now possible to recognize the broad outlines of what might be called the neoclassical theory. In the discussion which follows we will group the subjects under four main headings corresponding with four basic intuitions. The first topic is efficient markets, which was also the first area of finance that matured into a science. Next come the twinned subjects of return and risk. This leads naturally into option pricing theory and the central intuition of pricing close substitutes by the absence of arbitrage.

The principle of no arbitrage is used to tie together the major subfields of finance. The fourth section looks at corporate finance from its well-developed form as a consequence of no arbitrage to its current probings. A short conclusion ends the entry.

Efficient Markets

The word efficient is too useful to be monopolized by a single meaning in economics. As a consequence, it has a variety of related but distinct meanings. In neoclassical equilibrium theory efficiency refers to Pareto efficiency. A system is Pareto efficient if there is no way to improve the well being of any one individual without making someone worse off. Productive efficiency is an implication of Pareto efficiency. An economy is productively efficient if it is not possible to produce more of any one good or service without lowering the output of some other.

In finance the word efficiency has taken on quite a different meaning. A capital market is said to be (informationally) efficient if it utilizes all of the available information in setting the prices of assets. This definition is purposely vague and it is designed more to capture an intuition than to state a formal mathematical result. The basic intuition of efficient markets is that individual traders process the information that is available to them and take positions in assets in response to their information as well as to their personal situations. The market price aggregates this diverse information and in that sense it 'reflects' the available information.

The relation between the definitions of efficiency is not obvious, but it is not unreasonable to think of the efficient markets definition of finance as being a requirement for a competitive economy to be Pareto efficient. Presumably, if prices did not depend on the information available to the economy, then it would only be by accident that they could be set in such a way as to guarantee a Pareto efficient allocation (at least with respect to the commonly held information).

If the capital market is competitive and efficient, then neoclassical reasoning implies that the return that an investor expects to get on an investment in an asset will be equal to the opportunity cost of using the funds. The exact specification of the opportunity cost is the subject of the section on risk and return, but for the moment we can observe that investing in risky assets should carry with it some additional measure of return beyond that on riskless assets to induce risk averse investors to part with their funds. For now we will defer the measurement of this risk premium, and simply represent the opportunity cost by the letter 'r'.

In much of the early empirical work on efficient markets no attempt was made to measure risk premia, and the opportunity cost of investing was set equal to the riskless rate of interest. This can be justified either by assuming that there are risk neutral investors who are indifferent to risk (or, as we shall see, by assuming that the asset's risk is diversified away in large portfolios). Whatever the rationale, to focus on the topic of efficient markets rather than on the pricing of risk, we will let r be the riskless interest rate.

If R_t denotes the total return on the asset – capital gains as well as payouts – over a holding period from t to $t+1$, then the efficient markets hypothesis (EMH) asserts that

$$E(R_t|I_t) = (1 + r_t), \tag{1}$$

where E is the expectation taken with respect to a given information set I_t, that is available at time t (and that includes r_t). An alternative formulation of the basic EMH equation is in terms of prices. For an asset with no payouts, since

$$R_t \equiv p_{t+1}/p_t,$$

we can rewrite (1) as

$$E(p_{t+1}|I_t) = (1 + r_t)p_t, \tag{2}$$

or, equivalently, discounted prices must follow the martingale,

$$\frac{1}{(1 + r_t)}E(p_{t+1}|I_t) = p_t.$$

The EMH is given empirical content by specifying the information set that issued to determine

prices. Harry Roberts (1967) first coined the terms which have come to describe the categories of information sets and, concomitantly, of efficient market theories that are employed in empirical work. Fama (1970) subsequently articulated them in the form which we now use. These categories describe a hierarchy of nested information sets. As we go up the hierarchy from the smallest to the biggest set (i.e. from coarser to finer partitions) we are requiring efficiency with respect to increasing amounts of information. At the far end of the spectrum is strong-form efficiency. Strong-form efficiency asserts that the information set, I_t, used by the market to set prices at each date t contains all of the available information that could possibly be relevant to pricing the asset. Not only is all publicly available information embodied in the price, but all privately held information as well.

A substantial notch down from strong-form efficiency is semistrong-form efficiency. A market is efficient in the semistrong sense if it uses all of the publicly available information. The important distinction is that the information set, I_t, is not assumed to include privately held information, i.e. information that has not been made public. Making this distinction precise is possible in formal models but categorizing information as publicly available or not can be subjective. Presumably, accounting information such as the income statements and the balance sheets of the firm is publicly available, as is any other information that the government mandates should be released such as the stock holdings of the top executives in the firm. Presumably, too, the true but unrevealed intention of a major stockholder would fall into the category of private information. In between these extremes is a large grey area.

The tendency in the empirical literature has been to take a purist's view of semistrong efficiency, and to adopt the position that if the information was in the public domain then it was available to the public and should be reflected in prices. This ignores the cost of acquiring the information, but the intuitive justification for this position is that the costs of acquiring such public information are small compared to the potential rewards. Thus, while the government mandated and publicly reported trades of the top executives require a bit more effort to obtain in a timely fashion than some average of their past holdings, such trades, when reported, would fall squarely within the realm of publicly available information under the semistrong version of the EMH.

If the asset is traded on an organized exchange, then of all the information that is clearly available to the public, none is as accessible and cheap as its past price history. At the bottom of the ladder in the efficiency hierarchy, weak-form efficiency requires only that the current and past price history be incorporated in the information set. If there is empirical validity to the EMH then, at the very least, the market for an asset should be weak-form efficient, that is, efficient with respect to its own past price history.

Empirical Testing

The empirical implications of efficiency with respect to a particular information set are that the current price of the asset embodies all of the information in that set.

Since the categories of information sets are nested, rejection of any one type, say, weak-form efficiency, implies the rejection of all stronger forms.

For example, according to weak-form efficiency, the current price of an asset embodies all of the information contained in the past price history. This implies that,

$$E(R_t | R_{t-1}, R_{t-2}, \ldots) = (1 + r_2), \qquad (3)$$

or, in price terms,

$$E(p_{t+1} | p_t, p_{t-1}, \ldots) = (1 + r_t)p_t.$$

The most dramatic consequence of the EMH and certainly the one that receives the most attention from the public, is that it denies the possibility of successful trading schemes. If, for example, the market is weak-form efficient, then an investor who makes use of the 'technical' information of past prices can only expect to receive a return of the opportunity cost $(1 + r_t)$. No amount of clever manipulation of the past information can improve this result.

As a test of weak-form efficiency, then, we could test (although not as a simple regression) the null hypothesis that

$$H_0 : E\left(p_{t+1} | p_t, p_{t-1}\right) = \beta_0 + \beta_1 p_t + \beta_2 p_{t-1}, \quad (4)$$

where

$$\beta_0 = 0 \quad \beta_1 = (1 + r_t).$$

and

$$\beta_2 = 0.$$

The important feature of this hypothesis is that it tells what information does *not* play a role (given r_t), namely the lagged price, $p_{t—1}$. If the coefficient $\beta 2$ should prove to be statistically significant, then this would constitute a rejection of the weak-form EMH.

The other empirical implication of the EMH that is often cited as a defining characteristic is that an efficient price series should 'move randomly'. The precise meaning of this in our context is that price changes should be serially uncorrelated.

Consider the serial covariance between two adjacent rates of return,

$$\begin{aligned} &\operatorname{cov}(R_{t+1}, R_t) \\ &\quad \equiv E([R_{t+1} - E(R_{t+1})][R_t - E(R_t)]). \\ &\quad = E(R_{t+1}[R_t - E(R_t)]) \\ &\quad = E(E(R_{t+1}|R_t)[R_t - E(R_t)]) \end{aligned} \quad (5)$$

In Eq. 5, since we have not specified the information set with respect to which the expectations are to be taken, they are unconditional expectations. Under weak- form efficiency, the information set will contain the past rates of return. Suppose that the (expected) opportunity cost, e.g. the interest rater r, independent of past returns on the asset or that changes are of a second order of magnitude. This would occur, for example, if we held r_t constant at r. In such a case, since weak-form efficiency implies that I_{t+1} contains R_t, we have

$$\begin{aligned} E(R_{t+1}|R_t) &= E[E(R_{t+1}|I_{t+1})|R_t] \\ &= E[(1 + r_{t+1})|R_t] \\ &= E(1 + r_{t+1}), \end{aligned} \quad (6)$$

the unconditional expectation of next period's opportunity cost. Putting (5) and (6) together yields,

$$\operatorname{cov}\left(R_{t+1}, R_t\right) = E(1 + r_{t+1})E[R_t - E(R_t)] = 0. \quad (7)$$

which is to say that rates of return are serially uncorrelated.

Tests of the EMH are legion and by and large they have been supportive. The early tests were essentially tests of the inability of trading schemes or of the random walk nature of prices, which implies that actual rates of return are serially uncorrelated. While the EMH does not imply that prices follow a random walk, such a price process is consistent with market efficiency. Alternatively, unable to specify closely the opportunity cost, some of the early tests took refuge in the view that it must be positive, which leads to a submartingale model for prices,

$$E\left(p_{t+1} | \mathrm{I}_t\right) \geq p_t. \quad (8)$$

The lack of a specification of the opportunity cost characterizes the early tests (see Cowles (1933), Granger and Morgenstern (1962), Cootner (1964) and see Roll's (1984) study of the orange juice futures market for a modern example of such a test). Following Fama (1970), the literature shifted to a concern for specifying the opportunity cost and, in this sense, empirical tests became joint tests of the EMH and of the correct specification of the opportunity cost and its attendant theory.

In terms of the information hierarchy, the general message that emerged from the testing is that the market does appear to be consistent with weak-form efficiency. Tests of stronger forms of efficiency, though, have produced mixed results. Fama et al. (1969) introduced a new methodology to test semistrong efficiency and applied it to stock splits. They observed that the residuals from a

F

simple regression of a stock's returns on a market index would measure the portion of the return that was not attributable to market movements. By adding there siduals over a period of time, the resulting cumulative residual measures the total return over that period that is attributable to non-market movements. If a stock splits, say, 2 for 1, then under semistrong efficiency its price should split in proportion. i.e., halve for a 2 for 1 split. Using this 'event study' approach, Fama et al. verified that stock split data was consistent with semistrong efficiency. The event study methodology they introduced and the use of cumulative residuals (averaged over firms) has become the standard method for examining the impact of information on stock returns.

By contrast with their supportive findings, Jaffé (1974), for example, found that a rule based on the publicly released information about insider trades produced abnormal returns. These results and others like them (see the section on *Risk and Return* below) have been much debated and no final verdict on the matter is likely.

Recently a more interesting empirical challenge to the EMH has come from a different tack. Shiller (1981), has argued that the traditional statistical tests that have been employed are too weak to examine the EMH properly and, moreover, that they are misfocused. Shiller adopts the intuitive perspective that if stock prices are discounted expected dividends, then they ought not to vary over time as much as actual dividends. He argues that since the price is an expectation of the dividends and future price, what actually occurs will be this expectation plus the error in the forecast and should be more variable than the price. This leads him to formulate statistical tests of the EMH based on the volatility of stock prices which are claimed to be more powerful than the traditional (regression based) tests.

An alternative view has been taken by critics of this perspective, notably Kleidon (1986), Flavin (1983), and Marsh and Merton (1986). These critics have taken issue with Shiller's specification of the statistical tests of volatility and, more importantly, with his basic intuition. In particular, they contend that the single realization of dividends and prices that is observed is only one drawing from all of the random possibilities and that the price is based on the expectation taken over all of these possibilities. A little bit of information, then, can have an important influence on the current price. Furthermore, they argue that when the smoothing of dividends and the finite time horizon of the data samples are taken into account, volatility tests do not reject the EMH. The testing of the EMH is taking a new direction because of this work, but, at present, the results are still mixed.

Less cosmic in scope, but perhaps more worrisome is the discovery by French and Roll (1985) that the variance per unit time of market returns over periods when the market is closed (for example, from Tuesday's close to Thursday's close when the market was closed on Wednesday because of a backlog of paperwork) is many times smaller than when it is open. It is difficult to reconcile this result with the requirement that prices reflect information about the cash flows of the assets, unless the generation of fundamental information slows dramatically when the market closes – no matter why it is closed.

Theoretical Formulations

The attempts to formalize the EMH as a consistent, analytical economic theory have met with less success than the empirical tests of the hypothesis. The theory can be broken into two parts. The first part is neoclassical and is largely formulated in terms of models in which investors share a common information set. Such models focus on the intertemporal aspects of the theory and the changing shape of the information set.

It has long been recognized that a competitive economy with a single risk neutral investor would lead to the traditional efficient market theories with respect to the information set employed by that investor. More interestingly, Cox et al. (1985a) and Lucas (1978) have developed intertemporal rational expectations models each of which is consistent with certain versions of the efficient market theories.

There is, however, an important sense in which these models fail to capture the essential intuition of efficient markets. In informationally efficient markets, prices communicate information to

participants. Information possessed by one investor is communicated to another through the influence – however microscopic – that the first investor has on equilibrium prices. In models where investors have homogeneous information sets such information transfer is irrelevant.

A variety of attempts have been made to develop models of financial markets which can deal with such informational issues, but the task is formidable and a satisfactory resolution is not now in hand. This work parallels that of the neoclassical rational expectations view of macroeconomics. This is no accident since the rational expectations school of macroeconomics was very clearly influenced by the intuition of efficiency in finance. The original insight that prices reflect the available information lies at the heart of rational expectations macroeconomics. In this latter work aggregate prices, for example, not only provide the terms of trade for producers, they also inform producers about the aggregate state of production in the economy.

Perhaps the principal difficulty is that models with fully rational investors tend to break down. As investors apply the full scope of their analytical and reasoning talents, the result is an equilibrium in which they lack the incentive to engage in trade. (See Grossman 1976; Grossman and Stiglitz 1980; Diamond and Verrecchia 1981; Milgrom and Stokey 1982; Admati 1985.) The only way out of this bind seems to be to add a discomforting element of irrationality – or an alternative motive for trade from an equilibrium, such as insurance – to the model.

To understand this point, consider a risk-averse individual trading in a market where he or she receives information signals about the ultimate value of the asset being traded and where it is common knowledge that all investors are in the same position. That is not to say that all investors have the same information, rather, it only means that they all begin with the same information, have the same view of the world (Bayesian priors), and then receive signals from the same sort of information generating mechanism. In such a market, the offer to trade on the part of any one investor communicates information to other investors. In particular, it tells them that the individual, based upon his or her information, will be improved by the trade. If all investors are rational they will all feel similarly bettered by trade. But, if the market had been in an equilibrium prior to the receipt of new information, and if it is common knowledge that trade balances, then in the new equilibrium not all of them can be improved. This contradiction can only be resolved by having no further trade upon the receipt of information.

To put the matter in an equivalent form, consider an investor who possesses some special information. Presumably, it is by trading that this information is incorporated into the market price. The above argument implies that the mere announcement of a wish to trade results in a change in prices with no profits for the investor since none will trade at the original prices. If information is costly to acquire and impossible to profit from, then why bother? In other words, if the price reflects the available information possessed by the individual participants, then why gather information if one only needs to look at the price?

The resolution of this dilemma can take many forms, and research will proceed by altering the assumptions that lead to this result. For example, we can drop the assumption about a common prior and let investors come to the markets with different a priori beliefs. We could also drop the assumption that all investors are perfectly rational and introduce 'noisy' traders. Lastly, we could drop efficiency and complete markets or integrate insurance motives in other ways.

All of these approaches are being explored but we must leave this discussion with the theory that underlies the incorporation of asymmetric information into securities prices in an unsettled state. The traditional theory that prices reflect the available information is well understood with a representative individual. The theory with asymmetric information is not well understood at all. In short, the exact mechanism by which prices incorporate information is still a mystery and an attendant theory of volume is simply missing.

To conclude, the efficient market paradigm is the backbone of much of financial research and it

continues to guide a large body of theoretical and empirical work. Its usefulness is beyond question, but its fine structure is not. In a sense, like much of economics, it remains a central intuition whose analytical representations seem less compelling than the insight itself. This presents more of a problem for theory than for empirical work, but the empirical side is also not without challenge. Although the evidence in support of the efficiency of capital markets is widespread, troublesome pockets of anomalies are growing and the power of the traditional methodology to test the theory is being seriously questioned. Nevertheless, there is currently no competitor for the basic intuition of efficient markets and few insights have proven as fruitful.

Risk and Return

The theory of efficient markets leads inexorably to the second central intuition in finance, the trade-off between risk and return. It has long been recognized that risk-averse investors require additional return to bear additional risk. Indeed, this insight goes back to the earliest writings on gambling and it is as much a definition of risk aversion as it is a description of risk-averse behaviour. The contribution made by finance has been to translate this observation into a body of intuition, theory, and empirics on the workings of the capital markets.

The intuition that in a competitive market higher return is accompanied by higher risk owes at least as much to Calvin as it does to Adam Smith, but, in large part the development of capital market theory has been an attempt to explain risk premia, the difference between expected returns and the riskless interest rate. The foundations for the models that would first explain risk premia and that would become the workhorses of financial asset pricing theories were laid by Hicks (1946), Markowitz (1959), and Tobin (1958). These authors developed a rigorous micro- model of individual behaviour in a 'mean variance' world where investment portfolios were evaluated in terms of their mean returns and the total variance of their returns. They justified focusing on these two distributional characteristics by assuming either that investors had quadratic von Neumann-Morgenstern utility functions or that asset returns were normally distributed. In such a world, investors would choose mean variance efficient portfolios, i.e., portfolios with the highest mean return for a given level of variance. This observation reduced the study of portfolio choice to the analysis of the properties of the mean variance efficient set. Building on their work, Sharpe (1964), Lintner (1965), and Mossin (1966), all came to the fundamental insight that this micromodel could be aggregated into a simple model of equilibrium in the capital markets, the capital asset pricing model or CAPM.

The Mean Variance Capital Asset Pricing Model (CAPM)

In neoclassical equilibrium models, an investor evaluates an asset in terms of its marginal contribution to his or her portfolio. The decision to alter the proportion of the portfolio invested in an asset will depend on whether the cost of doing so in terms of risk is greater or less than the benefit in expected return. An individual in a personal equilibrium will find the cost at the margin equal to the benefit.

We will assume that a unit addition of an asset to the portfolio can be financed at an interest rate of r. In a mean variance model the net benefit of adding an asset to a portfolio is the additional expected return it brings, E, less the cost of financing it. Such a change, Δx, will augment the expected return on the portfolio, E_p, by the risk premium of the asset, i.e. by the difference between the expected return on the asset, E_i, and the cost of the financing r,

$$\Delta E_p = (E_i - r)\Delta x. \tag{9}$$

The marginal cost, in terms of risk, of an increase in the holding of an asset is the addition to the total variance of the portfolio occasioned by an increase in the holding of the asset. To compute this increase, let v denote the variance of returns on the current portfolio, let var (i) stand for the variance of asset i's returns, let cov(i, p) denote the covariance between the return of asset i and that of

the portfolio, p, and let Δx be the addition in the holding of asset i.

The variance of the portfolio after adding $\Delta x$of asset i will be,

$$v + \Delta v = v + 2\Delta x \text{cov}(i,p) + (\Delta x)^2 \text{var}(i),$$

which means the change in the variance is given by

$$\Delta v = 2(\Delta x)\text{cov}(i,p) + (\Delta x)^2 \text{var}(i),$$

and for a small marginal change, Δx this approximates,

$$\Delta v \approx 2(\Delta x)\text{cov}(i,p).$$

The marginal rate of transformation between return and risk, then, is given by

$$\text{MRT} = \frac{\Delta E_p}{\Delta v} = \frac{(E_i - r)\Delta x}{2(\Delta x)\text{cov}(i,p)} = \frac{(E_i - r)}{2\text{cov}(i,p)}. \quad (10)$$

An investor will be in a personal equilibrium when this trade-off is equal to his or her personal marginal rate of substitution between return and risk. But, if the portfolio p is an optimal one for the investor then it must also have a trade-off between return and risk that is equal to the investor's marginal rate of substitution, and this permits us to use it as a benchmark. Consider, then, the alternative possibility of changing the portfolio position not by changing the amount of asset i being held, but rather by changing the amount of the entire portfolio p being held, again financing the change by an alteration in the holding of the riskless asset. This is equivalent to leveraging the portfolio of risky assets and altering the amount of the riskless asset so as to continue to satisfy the budget constraint. Such a change will produce a trade-off between return and risk exactly analogous to the one examined above.

$$\text{MRS} = \frac{E_p - r}{2\text{var}(p)}, \quad (11)$$

where we have written this as the marginal rate of substitution, MRS. Since in equilibrium all of the marginal rates of transformation must equal the common marginal rate of substitution, putting these two equations together we have,

$$E_i - r = (E_p - r)\beta_{ip}, \quad (12)$$

where

$$\beta_{ip} \equiv \frac{\text{cov}(i,p)}{\text{var}(p)}, \quad (13)$$

the regression coefficient of the returns of asset i on the returns of portfolio, p. Eq. 12 is the famous security market line equation, the SML. It describes the necessary and sufficient condition for a portfolio p to be mean variance efficient. It also provides a clear statement of the risk premium, asserting that it is proportional to the asset's beta, β_{ip}.

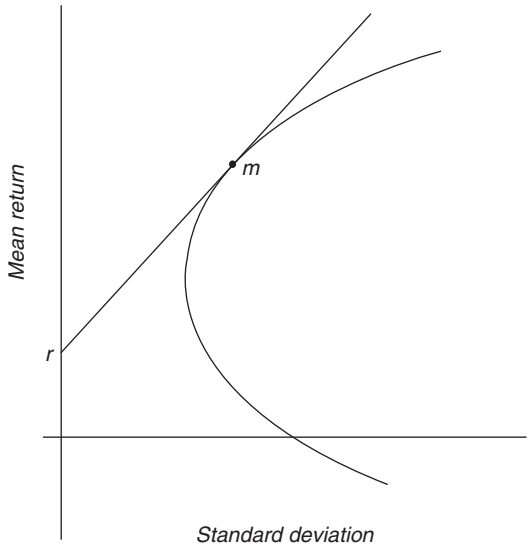
The insight of Sharpe, Lintner and Mossin was the observation that the SML and the mean variance analysis could be aggregated almost without change to a full equilibrium in the capital market. If we assume that all individuals have the same information and, therefore, see the same mean variance picture, then each individual's efficient portfolio will satisfy Eq. 12. Since the SML equation is linear in the portfolio holding, p, we can simply weight each individual's equation by the proportion of wealth that individual holds in equilibrium, and add up the individual SML's. The result will be an SML equation for the aggregate portfolio, m, that is the weighted average of the individual portfolios. In equilibrium, the weighted average of all of the individual portfolios, m, is the market portfolio, i.e., the portfolio of all assets held in proportion to their market valuation. In other words, each asset i, must lie on the SML with respect to the market,

$$E_i - r = (E_m - r)\beta_{im}, \quad (14)$$

which means that the market portfolio, m, is a mean variance efficient portfolio.

The geometry of the mean variance analysis is illustrated in Fig. 1. The set of mean variance

F



Finance, Fig. 1

efficient portfolios maps out a mean variance efficient frontier in the mean standard deviation space of Fig. 1. Each investor will pick some point on this frontier and that point will be associated with a mean variance efficient portfolio that is suitable for the investor's particular degree of risk aversion. All such portfolios will themselves be portfolios of just two assets: the riskless asset, r, and a common portfolio, p, of risky assets. This fortunate simplification of the individual portfolio optimization problem is referred to as two fund separation. It implies that the only role for individual preferences lies in choosing the appropriate combination of the risky portfolio, p, and the riskless asset, r. As a consequence, when we aggregate, the market risk premium, $(E_m - r)/\text{var}(m)$, will be an average of individual measures of risk aversion.

Black (1972) showed that two fund separation would still hold in the mean variance model even if there were no riskless asset. In such a case he found that an efficient portfolio orthogonal-the 'zero beta portfolio'-to the market portfolio could be found, and that all investors would be able to find their optimal portfolios as combinations of m and this zero beta portfolio. In the above development of the CAPM we can simply let r be the expected return on a zero beta portfolio.

The necessary and sufficient conditions on return distributions for them to have this two fund separation property – for any concave utility function – were established by Ross (1978a). Ross characterized the class of distribution whose efficient frontier, i.e. the set of portfolios that *some* investor would choose, was spanned by k funds, and showed that it extended beyond the normal distributions in the case of $k = 2$ fund separation. This work was extended by Chamberlain (1983), who found the class of distributions for which expected utility was a function of just mean and variance for any portfolio as well as for the efficient ones. Cass and Stiglitz (1970) found the conditions on investor utility functions for a similar property to hold regardless of assumptions on return distributions.

It follows immediately from two fund separation that the tangency portfolio, p, in Fig. 1 must be the market portfolio of risky assets since all investors hold all risky assets in the same proportions. If there is no net supply of the riskless asset then p must be the market portfolio, m, itself.

The central feature of the CAPM is the mean variance efficiency of the market portfolio and the emergence of the beta coefficient on the market portfolio as the determinant of the risk premium of an asset. Those features of an asset that contribute to its variance but do not affect its covariance with the market will not influence its pricing. Only beta matters for pricing; the idiosyncratic or unsystematic risk, i.e. that portion which is the residual in the regression of the asset's returns on the market's returns and is therefore orthogonal to the market, playing no role in pricing.

This produces some results that were at first viewed as counter-intuitive. The older view that the risk premium depended on the asset's variance was no longer appropriate, since if one asset had a higher covariance with the market than another, it would have a higher risk premium even if the total variance of its returns were lower. Even more surprising was the implication that a risky asset that was uncorrelated with the market would have no risk premium and would be expected to have the same rate of return as the riskless asset, and that assets that were inversely correlated with the

market would actually have expected returns of less than the riskless rate in equilibrium.

These results for the CAPM were supposedly explicated by the twin intuitions of diversification and systematic risk. There could be no premium for bearing unsystematic risk since a large and well diversified portfolio (i.e. one whose asset proportions are not concentrated in a small subset) would eliminate it – presumably by the law of large numbers. This would leave only systematic risk in any optimal portfolio and since this risk cannot be eliminated by diversification, it has to have a risk premium to entice risk averse investors to hold risky assets. From this perspective it becomes clear why an asset that is uncorrelated with the market bears no risk premium. One that is inversely correlated with the market actually offers some insurance against the all pervasive systematic risk and, therefore, there must be a payment for the insurance in the form of a negative risk premium.

There is nothing wrong with this intuition, but it does not fit the CAPM very well. The residuals from the regression of asset returns on the market portfolios are orthogonal to the market, but they could be highly correlated with each other. In fact, they are linearly dependent since when they are weighted by the market proportions they sum to zero. This means the law of large numbers cannot be used to insure that large portfolios of residuals other than the market portfolio will be negligible. But, if that is the case, then the residuals could capture systematic risks not reflected in the market portfolio.

The CAPM was the genesis for countless empirical tests (see, e.g., Black et al. 1972 and Fama and MacBeth 1973). The latter paper developed the most widely used technique. The general structure of these tests was the combination of the efficient market hypothesis with time series and cross section econometrics. Typically some index of the market, such as the value weighted combination of all stocks would be chosen and a sample of firms would be tested to see if their excess returns, $E-r$, were 'explained' in cross-section by their betas on the index, i.e., whether the SML was rejected.

Roll (1977b, 1978) put a stop to this indiscriminate testing by calling into question precisely what was being tested. Roll's critique had two parts. First, he argued that the tests were of very low power and probably could not detect departures from mean variance efficiency. His central point, though, began by noting that tests of the CAPM were tests of the implications of the statement that the *entire* market portfolio was mean variance efficient, and were not simply tests of the efficiency of some limited index such as could be formed from the stock market. The essential role played by the market portfolio in the CAPM had been stressed by others; Ross (1977b) had shown the equivalence between the CAPM and the mean variance efficiency of the market portfolio. (Ross (1976a) had also shown that in the absence of arbitrage there was always some efficient portfolio.) Roll went beyond this simple observation, though, by stressing the essential point that the market portfolio is unmeasurable. This called into question the entire cottage industry of testing the CAPM and all of the uses to which the theory had been put, such as performance measurement.

Inter-Temporal Models

In the aftermath of Roll's critique, attention was turned to alternative models of asset pricing and the intertemporal nature of the theory became more important. Two separate strands of development can be traced. One essentially followed the lines of the CAPM and developed the intertemporal versions of it, the ICAPM. Merton (1973a) pioneered in this. Using continuous time diffusion analysis, Merton showed that the CAPM could be generalized to an intertemporal setting. Most interestingly, though, he demonstrated that, if the economic environment was described by a finite dimensional vector of state variables, x, and if asset prices were exogenously specified random variables, then a version of the SML would hold at all moments of time with the addition to the risk premium of a linear combination of the betas between the assets' returns and each of the state variables, x_i.

Ross (1975) developed a similar inter-temporal extension of the CAPM, but Ross's model simplified preferences in order to close the model with an inter-temporal rationality constraint and to

study equilibrium price dynamics. Along the lines being developed in the modern literature on macroeconomics, inter-temporal rationality and the efficient market theory required that the distribution of prices be determined endogenously. A discrete time Markov model with this feature was presented in Lucas (1978) and a full rational expectations general equilibrium in continuous time was developed in Cox et al. (1985a).

Cox et al. (1985b) applied their model to analyse and resolve some long-standing questions in the theory of the term structure of interest rates. The theory of the term structure is one of the most important subfields of finance, and the bond markets were one of the first areas where the EMH was applied. In an efficient market, ignoring risk aversion, forward rates should be (unbiased) predictors of future spot rates and many early theories and tests of the EMH were formulated to examine this proposition (see e.g. Malkiel 1966). Roll (1970) integrated the EMH with the CAPM and used the resulting framework to examine empirically liquidity premia in the bond markets; the work of Cox et al. (1985b) can be considered as the logical extension of his analysis to a rational inter-temporal setting.

Merton's model was simplified markedly by Breeden (1979), who showed that, if investors had intertemporally additive utility functions, then Merton's ICAPM and its version of the SML could be collapsed back into a single beta model, the consumption beta model, with all assets being priced, that is, having their risk premiums determined, by their covariance with aggregate consumption (see also Rubinstein (1976)). If we think of returns as relative prices between wealth today and in future states of nature, then optimizing individuals will set their marginal rates of substitution between consumption today and in future states equal to the rates of return. With continuous asset prices and additive utility functions, indirect utility functions are locally quadratic in consumption and this implies that consumption plays the role of wealth in the static CAPM. This work led to a variety of attempts to measure the ability of betas on aggregate consumption to explain risk premia (see e.g. Hansen and Singleton 1983).

Arbitrage Pricing Theory (APT)

A separable but related strand of theory is the arbitrage pricing theory (APT) (see e.g. Ross 1976a, b). The CAPM and the consumption beta model share the common feature that they explain pricing in terms of endogenous market aggregates, the market portfolio, and aggregate consumption, respectively. The APT takes a different tack.

The intuition of the CAPM (or of the Consumption Beta model) is that idiosyncratic risk can be diversified away, leaving only the systematic risk to be priced. Idiosyncratic risk, though, is defined with reference to the market portfolio as the residual from a regression of returns on the market portfolio's returns. Since no further assumptions are made about the residuals, contrary to intuition a large diversified portfolio that differs from the market portfolio will not in general have insignificant residual risk. The exception is the market portfolio, but then the intuition that diversification leads to pricing by the market portfolio is circular at best.

The APT addresses this issue by assuming directly a return structure in which the systematic and idiosyncratic components of returns are defined a priori. Asset returns are assumed to satisfy a linear factor model,

$$R_i = E_i + \sum_j \beta_{ij} f_j + \varepsilon_i, \quad i = 1, \ldots, n, \qquad (15)$$

where E_i, is the expected return, f_j is a demeaned exogenous factor influencing each asset i through its beta on the factor, β_{ij}, and ε_i is an idiosyncratic mean zero term assumed to be sufficiently uncorrelated across assets that it is negligible in large portfolios. An implication of the factor structure is that the ε terms become negligible in large well diversified portfolios and, therefore, such portfolios approximately follow an exact factor structure,

$$R_i \approx E_i + \sum_j \beta_{ij} f_j, \qquad (16)$$

where i now denotes the ith well diversified portfolio. In an Arrow–Debreu state space framework,

Eq. 16 can be interpreted as a restriction on the rank of the state-space tableaux.

An exact factor structure implies that there will be arbitrage unless the expected return on each portfolio is equal to a linear combination of the beta coefficients,

$$E_i - r = \sum_j \lambda_j \beta_{ij}, \tag{17}$$

where λ_j is the risk premium associated with the jth factor, f_j. This equation is the APT version of the SML in the CAPM.

The APT is consistent with a wide variety of equilibrium models (including the CAPM if there is a factor structure) and it has been the object of much theoretical and empirical attention. In a sense, the APT can be thought of as a snapshot of any intertemporal model in which the factors represent innovations in the underlying state variables. This means that a rejection of the APT would imply a fairly wide ranging rejection of attempts to model asset markets with a finite set of state variables.

The original theoretical development of the APT (Ross 1976a, b) showed formally that, if preferences are continuous in the quadratic mean, then the returns on a sequence of portfolios which require no wealth cannot converge to a positive return with a zero variance. This, in turn, implies that the sum of squared deviations from exact APT pricing is bounded above. These results were simplified by Huberman (1982) and extended by Ingersoll (1984) and Chamberlain and Rothschild (1983), all of whom side-stepped the issue of preferences by simply assuming that there could be no sequences converging to an arbitrage situation of a positive return with no variance. By contrast, Dybvig (1983) makes assumptions on preferences and aggregate supply to obtain a tight bound on pricing. His simple order of magnitude calculation is evidence that the pricing error is too small to be of practical significance.

By modelling the capital market explicitly as responding to innovations in exogenous variables, the APT is immediately inter-temporally rational. By contrast with the CAPM and the Consumption Beta models which price assets in terms of their relation with a potentially observable and endogenous market aggregate (wealth for the CAPM and consumption for the Consumption Beta models), the APT factors are exogenous, but unspecified. Much empirical work is now under way to determine a suitable set of factors for representing systematic risk in a factor structure and to examine if they price assets successfully. (For example, see Roll and Ross 1980; Brown and Weinstein 1983; Chen et al. 1986.)

The lack of an a priori specification for the factors has been the focus of criticism of the testability of the APT by Shanken (1982). Shanken argues that, since the factors are not pre-specified, the intuitive derivation of the APT given above can be used to verify APT pricing falsely even when it does not hold, and that to prevent this some equilibrium model, such as that proposed by Connor (1984), must be used. Shanken emphasizes that his critique applies not to the theory of the APT but rather to the way in which it has been tested. Dybvig and Ross (1985) dispute his arguments, stressing that Shanken wants to test the theory including its assumptions and approximations rather than take the positive approach of testing the model's conclusions.

Empirical Testing of Asset Pricing Models

Since Roll's critique, the methodology for testing asset pricing models has changed. There has been a retreat from testing a model per se to an explicit view that what is being tested is not the CAPM, for example, but rather whether the particular index being used for pricing is mean variance efficient. This change of focus has led to a more formal approach to the statistics of testing. Ross (1980) developed the maximum likelihood test statistic for the efficiency of a given portfolio and pointed out the analogy between this and the mean variance geometry, and Gibbons (1982) showed that the test of efficiency could be conducted by the use of seemingly unrelated regressions. These results have been extended by others. For example, Kandel (1984) and Jobson and Korkie (1982) and Gibbons et al. (1986) have developed and exploited an exact small sample test of the efficiency of a given index in the

presence of a riskless asset. Similar tests of the APT have not yet been developed, and to date much of the testing of the APT has focused on comparisons between the APT and pricing using the value weighted index (see e.g. Chen et al. 1986).

The most important empirical finding in asset pricing, though, has been the discovery of a wide array of phenomena that appear to be inconsistent with nearly any neoclassical model. Consider, first, the secular effects. Asset returns fall, on average, over the weekend and rise during the week (see French 1980). Similarly, it has been found that asset returns behave differently in the first half of the month than they do in the second. The most attention, though, has been lavished on the 'small firm effect'. It appears that the average returns on small firms exceed those on large firms no matter what theory of asset pricing is used to correct for differences in the risk premium between these two categories of assets. Furthermore, the bulk of the return difference is concentrated in the first few days of January. Indeed, on average, returns in January appear to be abnormally large for all stocks (see, for example, Keim 1983 or Roll 1981, 1983).

Potentially these sorts of anomalies can be explained by secular changes in risk premia – perhaps due to secular patterns in the release of information – but their persistence and magnitude make them serious challenges to all the asset pricing models. When evidence of this sort appears difficult to explain by any pricing model it calls into question the efficient market hypothesis itself. Tests of an asset pricing model are usually joint tests of both market efficiency and the pricing model; rejecting a wide enough range of such models is tantamount to rejecting efficiency itself.

Substitution and Arbitrage: Option Pricing

The APT is the child of one of the central intuitions of finance, namely, that close substitutes have the same price. This intuition reached fruition in the path breaking paper by Black and Scholes (1973) on option pricing. Since then the theory has found myriad applications and has been significantly extended; see, for example, Merton (1973b), Cox and Ross (1976a, b, c), Rubinstein (1976), Ingersoll (1977) and Cox et al. (1979, 1985a). The Black-Scholes model employed stochastic calculus, but a simpler framework for option pricing was presented by Cox et al. (1979) that retained its essential features and was more flexible for computational purposes. We will briefly outline this binomial approach and show its connections to the major theoretical features of option pricing.

The Binomial Model

The binomial model begins with the assumption that the price of a stock, S, follows a proportional geometric process:

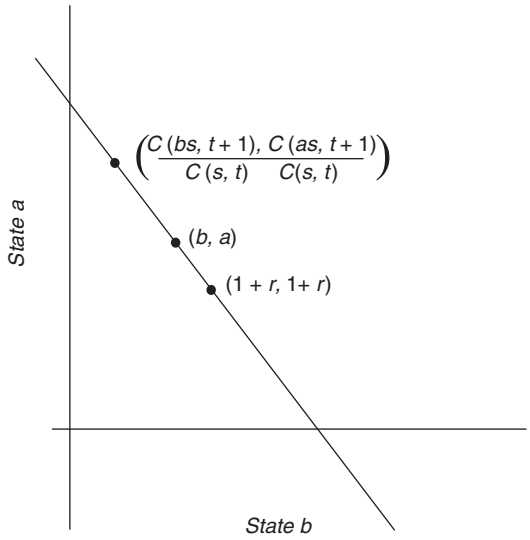
$$S(t+1) = \begin{cases} aS(t) & \text{with probability } \pi \\ bS(t) & \text{with probability } 1-\pi. \end{cases} \tag{18}$$

In addition to the stock there is also a riskless bond with a return of $1 + r$. The basic problem of option pricing theory is to determine the value of a derivative security, i. e., a security whose payoff depends only upon the value of an underlying primitive security, the stock in this case.

Let $C(s, t)$ denote the value of the derivative security as a function of the price of the stock and the time, t. Since its value depends only upon the movement of the stock – a result that is sometimes derived as a function of other attributes such as its value at the end of some period – it will also follow a binomial process:

$$C(S,t+1) = \begin{cases} C(aS,t) & \text{with probability } \pi \\ C(bS,t) & \text{with probability } 1-\pi. \end{cases} \tag{19}$$

The time $t + 1$ values are illustrated in Fig. 2. At any moment of time the information structure branches into relevant states, state a and state b, defined by whether the stock goes up by a or b. As the figure is drawn, $a > 1 + r > b$, and clearly $1 + r$ must lie between a and b to prevent the stock or



Finance, Fig. 2

the bond dominating. At this point there are two separate approaches to the analysis. The first is in the spirit of the original Black-Scholes model.

Suppose that at time t we form a portfolio of the riskless bond and the stock with α dollars invested in the stock and $1-\alpha$ dollars invested in the bond. We will choose the investment proportion so that the return on the portfolio coincides with the return on the derivative security in state b. This means choosing α so that

$$\frac{C(bS,t+1)}{C(S,t)} = \alpha b + (1-\alpha)(1+r), \tag{20}$$

which implies that

$$\alpha = \frac{(1+r) - C(bS,t+1)/C(S,t)}{(1+r)-b}. \tag{21}$$

But, since the portfolio's return matches that of the derivative security in state b, it must also match it in state a. If it did not, then either the portfolio or the derivative security would dominate the other, which would be an arbitrage opportunity. In other words, we must have,

$$\frac{C(bS,t+1)}{C(S,t)} = \alpha a + (1-\alpha)(1+r). \tag{22}$$

Putting these two equations together produces a difference equation which is satisfied by the value of the derivative security,

$$\pi^* C(aS,t+1) + (1-\pi^*)C(bS,t+1) - (1+r)C(S,t) = 0, \tag{23}$$

where

$$\pi^* \equiv \frac{(1+r)-b}{a-b}. \tag{24}$$

Perhaps the most remarkable feature of this equation is that it does not involve the original probabilities for the process, π, but rather is a function of what are called the martingale probabilities, π^*.

To solve this difference equation for the value, C, of a particular derivative security we would need only to append the contractual boundary conditions that define it. For example, a European call option is specified to have the value max (S–E, 0), at a specified future date, T, where E is its exercise price. Such an option gives the holder the right – but not the obligation – to buy the stock for E at time T. The dual security is a European put option which gives the holder the right, but again not the obligation, to sell the stock for E, at time T. The problem is more difficult if the derivative security is of the American variety which means that the holder may exercise it any time up to and including the maturity date T and need not wait until T.

Soon after the Black–Scholes paper, Merton (1973b) examined a variety of option contracts and showed how extensive was the range of the technique. Notably, Merton was able to derive a number of qualitative results on option pricing that were relatively independent of the particular process being modelled. For example, he showed that an American call option on a stock that pays no dividends will never be exercised before its maturity date and, therefore, will have the same value as a similar European call. He also demonstrated that put/call parity, i.e. the equivalence between the positions of holding the stock and a put option and holding a bond and a call option, was not generally valid for American options. Ross (1976c) showed that the literature's emphasis on

F

puts and calls was not misplaced since any derivative security could be composed of puts and calls.

A second approach to the valuation problem in our simple example illuminates why the original probabilities played no role in the analysis. Figure 2 displays what is essentially a two-state Arrow-Debreu model. In such a model if there are two pure contingent claims contracts paying one dollar in each state, then all securities can be valued as a function of their values, q_a and q_b. It follows, then, that any two securities which are not linearly dependent will span the space just as two pure contingent claims would and they can be used to value all securities in the space.

In our example, the value of the bond is 1 and it must satisfy,

$$1 = q_a(1+r) + q_b(1+r), \tag{25}$$

and the value of the stock must satisfy,

$$S = q_a(aS) + q_b(bS),$$

or

$$1 = q_a a + q_b b. \tag{26}$$

Solving these two equations we can find the implicit values of the state contingent claims,

$$q_a = \frac{(1+r)-b}{(1+r)(a-b)},$$

and

$$q_b = \frac{a-(1+r)}{(1+r)(a-b)}. \tag{27}$$

Notice that these prices do not depend on the original probability, π, since they are derived from the values of the stock and the bond. Whatever influence the probability, π, has on values is already reflected in the returns on the stock and the bond, and the derivative security value will just be a function of the implicit state prices. Using these prices, it is readily verified that the difference equation for the value of the derivative security, Eq. 23, is the same as,

$$q_a C(aS, t+1) + q_b C(bS, t+1) = C(S, t). \tag{28}$$

Geometrically, this means that the point,

$$\{[C(bS, t+1)/C(S, t)], [C(aS, t+1)/C(S, t)]\},$$

we plot on the same line as the return points for the bond and the stock, $(1 + r, 1 + r)$ and (b, a). For a call option the point will be as drawn in Fig. 2 indicating that the call is more volatile than the stock.

Notice from (24) and (27) that

$$\pi^* = (1+r)q_a,$$

which means that the state space price can be interpreted as the discounted martingale probability. It is this interpretation that ties together the Cox and Ross (1976a) risk-neutral approach to solving option pricing problems and the general theory of the absence of arbitrage.

Cox and Ross (1976a) argued that since the difference equation that emerged for solving option pricing problems made no explicit use of any preference information, the resulting solution must also be independent of preferences. For example, then, the resulting solution must be the same as that which would obtain in a risk neutral world. In such a world, the state probabilities must be such that the expected returns on all assets are the same,

$$\pi^* a + (1-\pi^*)b = 1 + r,$$

where the solution for the probability, π*, is the same martingale probability defined above. For a European call option, then, the solution will be

$$\begin{aligned} C(S,t) &= \frac{1}{(1+r)^{T-t}} E^*[\max(S_T - E, 0)] \\ &= \frac{1}{(1+r)^{T-t}} \sum_{j \geq \ln\left[E/Sb^{-(T-t)}\right]/\ln(a/b)} \\ &(\pi^*)^j \times (1-\pi^*)^{T-t-j}\left(Sa^j b^{T-t-j} - E\right), \end{aligned} \tag{29}$$

where E^* is the expectation with respect to the martingale probabilities, π^* and $(1 - \pi^*)$. It is easily verified that (29) is the solution to the difference Eq. 23 subject to the boundary condition,

$$C(S,T) = \max(S - E, 0).$$

Contrast this formula with the original Black-Scholes formula for the value of a call option in a continuous time diffusion model,

$$C(S,t) = SN(d_1) - e^{-r(T-t)}N(d_2), \qquad (30)$$

where $N(\cdot)$ is the standard cumulative normal distribution function and,

$$d_1 \equiv \frac{\ln(S/E) + r(T-t) + \frac{1}{2}\sigma^2(T-t)}{\sigma\sqrt{(T-t)}},$$

and

$$d_2 \equiv d_1 - \sigma\sqrt{(T-t)}.$$

Equation 30 is the solution to the Black–Scholes option pricing differential equation,

$$\frac{1}{2}\sigma^2 S^2 C_{SS} + rSC_S - rC = -C_t, \qquad (31)$$

subject to the boundary condition,

$$C(S,T) = \max(S - E, 0).$$

The Black–Scholes differential Eq. 31 is derived from an analogous hedging argument to that for the binomial model, applied to the continuous lognormal stock process,

$$dS/S = \mu dt + \sigma dz$$

where z is a standard Brownian motion. In fact, as the time interval between jumps converges to zero and the jump sizes shrink appropriately, the binomial converges to the lognormal diffusion and its option pricing solution will converge to that for the lognormal diffusion. Notice, too, that in analogy with the binomial whose solution does not depend upon the state probabilities, the Black-Scholes option price (30) is independent of the expected return on the stock, μ.

The most interesting comparative statics result from these models is the observation that call or put option values increase with increasing variance, $\sigma 2$. This is a consequence of these options being convex functions of the terminal stock value, S_T (Cox and Ross 1976b).

The General Theory of Arbitrage

All of the above analysis can be tied together by the general theory of arbitrage. Under quite general conditions, it can be shown that the absence of arbitrage implies the existence of a linear pricing rule that values all of the assets (see e.g., Ross 1976a, 1978b; Harrison and Kreps 1979). In a static model with m states of nature, this means the existence of implicit state prices, q_j, such that $q_j > 0$, and such that any asset with payoffs of x_j in the states of nature will have the value,

$$p = \sum_j q_j x_j. \qquad (32)$$

The intertemporal extension of this result is most neatly displayed in terms of the martingale expectation used above. The absence of arbitrage now implies the existence of a martingale measure such that, with obvious notation,

$$p = E^*\left\{\exp\left[-\int_0^T r(s)ds\right]x_T\right\}.$$

This theory permits us to tie together not only the basic results of option pricing, but also our previous analysis of asset pricing models. For example, applying it to the exact factor model,

$$R_i = E_i + \sum_j \beta_{ij} f_j,$$

yields the APT,

$$1 = E^*(R_i) = E^*\left(E_i + \sum_j \beta_{ij} f_j\right)$$

$$= \frac{1}{(1+r)}\left[E_i + \sum_j \beta_{ij} E^*\left(f_j\right)\right].$$

or

$$E_i = (1 + r) + \sum_j \lambda_j \beta_{ij},$$

where

$$\lambda_j \equiv -E^*\left(f_j\right).$$

Similarly, in a mean variance framework the martingale analysis can be used to prove that there is always a portfolio whose covariances are proportional to the excess returns on each asset. In other words, the absence of arbitrage implies the existence of a mean variance efficient portfolio (see Ross 1976a; Chamberlain and Rothschild 1983).

Empirical Testing

Perhaps because the option pricing theory works so well, it has generated a surprisingly small empirical literature. Some early tests, for example, Black and Scholes (1973) and Galai (1977), focused on whether the models could be used to generate successful trading rules and found that any success was easily lost to transactions costs. Most interestingly, MacBeth and Merville (1979) found that the option formulas tended to underprice 'in the money' options and overprice 'out of the money' options, but Geske and Roll (1984) have argued that this effect disappears with a reformulation of the statistics.

Given a theory that works so well, the best empirical work will be to use it as a tool rather than to test it. Chiras and Manaster (1978), for example, show that implicit volatilities, i.e. variances computed by inverting the option formulas to obtain variance as a function of the quoted option price, have strong predictive power for explaining future realized stock variances. Patell and Wolfson (1979) use the implicit variances to examine whether stock prices are more volatile around earnings announcements.

These efforts should increase; options and option pricing theory give us an opportunity to measure directly the degree of anticipated uncertainty in the markets. Financial press terms such as 'investor confidence' take on new meaning when they can actually be measured.

This does not mean, however, that there are no important gaps in the theory. Perhaps of most importance, beyond numerical results (see, for example, Parkinson, 1977 or Brennan and Schwartz 1977), very little is known about most American options which expire in finite time. The American call option on a stock paying a dividend or the American put option are both easily solved in the infinite maturity case since the optimal exercise boundary is a fixed stock value independent of time (Merton 1973b; Cox and Ross 1976a, b). If dividends occur at discrete points, then if the call is exercised prematurely it will only be optimal to do so just prior to a dividend payment. This permits a recursive approach to the solution of this finite maturity option (see Roll 1977a; Geske 1979). But, with continuous payouts, surprisingly little is known about the exercise properties of either of these options in the American case.

Despite such gaps, when judged by its ability to explain the empirical data, option pricing theory is the most successful theory not only in finance, but in all of economics. It is now widely employed by the financial industry and its impact on economics has been far ranging. At a theoretical level, we now understand that option pricing theory is a manifestation of the force of arbitrage and that this is the same force that underlies much of neoclassical finance.

The Whole Is the Sum of the Parts – Corporate Finance

The use of arbitrage as a serious tool of analysis coincided with the beginning of the modern theory of corporate finance. In two seminal papers on the cost of capital, Modigliani and Miller (1958, 1963) argued that the overall cost of capital and, therefore, the value of the firm would be unaffected by its financing decision. Specifically, using arbitrage arguments, Modigliani and Miller showed that the debt/equity split would not alter a firm's value and they then argued that with the investment decision held constant, the dividend payout rate of the firm would also not affect that value. These two irrelevance propositions defined the study of corporate finance in much the same

way that Arrow's Impossibility Theorem defined social choice theory. At one and the same time they propounded an irreverent theory whose central feature was the irrelevance of the topic under study. This challenge, to weaken in a useful way the assumptions of their analysis, has guided research in this area ever since.

The Modigliani–Miller Analysis

Since the Modigliani-Miller (henceforth MM) irrelevance propositions are developed from the absence of arbitrage, they are quite robust to alternative specifications of the economic model. To derive the Modigliani-Miller propositions we will employ the no arbitrage theory above. Consider a firm which will liquidate all of its assets at the end of the current period, and let x denote the random liquidation value of the assets. Assume that the firm has debt outstanding with a face value of F and that the remainder of the value of the firm is owned by the stockholders who have the residual claim after the bondholders.

At the end of the period, if x is large enough the stockholders will receive $x - F$ and if x falls short of F they will receive nothing. Formally, then, the terminal payment to the stockholders is

$$\max(x - F, 0),$$

which will be recognized as the terminal payment on a call option. In other words – in a tribute to the ubiquitous nature of option pricing theory – the stockholders have a call option on the terminal value of the firm, x, with an exercise price equal to the face value of the debt, F. The bondholders can claim the entire assets if x is not sufficient to cover the promised payment of F, which means that they will receive,

$$\min(x, F)$$

The current value of the firm, V, is defined to be the value of all of the outstanding claims against its assets which in this case is the value of the stocks, S, and the bonds, B. Using the no arbitrage analysis, we find that (ignoring discounting),

$$\begin{aligned} V \equiv S + B &= E^*[\max(x - F, 0)] + E^*[\min(x, F)] \\ &= E^*[\max(x - F, 0) + \min(x, F)] \\ &= E^*(x), \end{aligned}$$

which is independent of the face value, F, of the debt and, therefore, independent of the relative amounts of debt and equity. This verifies the first of the MM irrelevance propositions.

To verify the irrelevance of value to the dividend payout, consider a firm about to pay a dividend, D. The current, pre-dividend, value of the stock is $p^-(D)$ and by the no arbitrage martingale analysis this is given by,

$$p^-(D) = E^*[D + p^+(D)] = D + E^*[p^+(D)],$$

where $p + (D)$ is the ex-dividend price. If the investment policy of the firm has been fixed, then the only impact that the current dividend payout can have on the stockholders is through its alteration of the cash in the firm. This means that changing the dividend to, say $D + \Delta D$, would necessitate a change in current assets of $-\Delta D$. From the first MM proposition the mode of financing this change in the dividend will be irrelevant to the determination of the firm's value and to simplify the analysis we will assume that it is financed by riskless debt. At an interest rate of r this would entail, say, a perpetual outflow from the firm of $r\Delta D$. Again applying the analysis and letting x_{t+s} be the cash flow at time $t+s$ given that a dividend of D is paid now, we have,

$$\begin{aligned} p^+(D+\Delta D) &= E^*\left[\int_0^\infty e^{-rst}(x_{t+s} - r\Delta D)\mathrm{d}s\right] \\ &= E^*\left(\int_0^\infty e^{-rs}x_{t+s}\mathrm{d}s\right) - E^*\left(\int_0^\infty e^{-rs}r\Delta D\mathrm{d}s\right) \\ &= p^+(D) - \Delta D. \end{aligned}$$

Thus, we have the irrelevance proposition,

$$\begin{aligned} p^-(D + \Delta D) &= E^*[(D + \Delta D) + p^+(D + \Delta D)] \\ &= D + \Delta D + E^*[p^+(D + \Delta D)] \\ &= D + \Delta D + E^*[p^+(D)] - \Delta D \\ &= D + E^*[p^+(D)] = p^-(D). \end{aligned}$$

F

The MM results were startling to those who had worked in corporate finance and had taken it for granted that the way in which a firm was financed affected its value. To understand the importance of the MM results for the most practical of problems; recall that the original impetus for the study of corporate finance was the determination of the firm's opportunity cost for investments, ρ. For a marginal investment, financed by the issuance of debt and equity, the cost of capital, ρ, also known as the weighted average cost of capital, WACC, would be the weighted average cost of the debt, r, and the cost of equity, k,

$$\rho = (S/V)k + (B/V)r, \tag{33}$$

(where we have ignored tax effects).

If debt is riskless, then r is the interest rate on such debt and k, the cost of equity, will be the return required by investors for the risk inherent in the stock. Presumably k could be found by appeal to one of the asset pricing models discussed above.

Now it is tempting to think, for example, that if $k > r$, then an increase in debt relative to equity will lower ρ. If this goes too far, debt will become risky and as r rises there will be a unique optimal debt/equity ratio, (B/S)0, that minimizes the cost of capital, ρ. This would be the discount rate to use for present value calculations and it would maximize the value of the firm. This was the traditional analysis of the leverage decision before MM.

By the MM theorem, though, value, V, is unaffected by leverage. This means that $\rho$is unaltered, since the total (expected) return to the stockholders and the bondholders, $Sk + Br$, is unaltered [see Eq. 33]. In terms of the WAAC, then, as the leverage (B/S) is increased by the substitution of debt for equity, the cost of equity changes.

$$k = \rho + (B/S)(\rho - r),$$

but not the WAAC.

Spanning Arguments

The efforts to elude these results and to develop a meaningful theory of corporate finance have taken many forms. First, it has been argued that the analysis itself contains a hidden and critical assumption, namely that the pricing operator is independent of the corporate financial structure. The alternative is that the change in the debt/equity decision, for example, will also change the span of the marketed assets in the economy and, consequently, the operator used for pricing will change. The simplest such example would be a single firm in a two-state world. If the firm is an all equity firm and if there are no other traded assets, then individuals cannot adjust their consumption across the states of nature and must split it according to the equity payoff. If this firm now issues debt the two securities will span the two states of nature and complete the market. This, in turn, will generally alter pricing in the economy.

While this argument has generated a large literature, the problem of the determination of the corporate financial structure and the value of the firm is primarily a microeconomic question and it is difficult to believe that it will be resolved or even illuminated by assuming that firms have some monopoly power that enables them to alter pricing in the capital markets. At the micro level the MM propositions are unlikely to be seriously affected by such general equilibrium arguments.

At the micro level, too, the intuition behind the MM propositions and its conclusions is so robust as to be daunting. Consider the following argument. According to MM there can be no optimal, i.e. value maximizing, financial structure since value is independent of structure. Suppose that there was an optimal, say, debt/equity ratio, $(B/S)^0$. Any departure from this target $(B/S)^0$, however, could not lower the firm's value since it would immediately afford an arbitrage opportunity to buy the total firm at its lowered value and refinance it in the optimal target propositions, $(B/S)^0$. (This somewhat facetious argument gets the point across, but it really means that we have not fully specified the rules of the game, e.g., who moves first, what happens when no one moves, etc.)

Signalling Models

A more promising route which formally exploits incomplete spanning, but does not argue that the pricing operator itself is altered by any one firm

changing its financial structure, makes use of the theory of asymmetric information and signalling (see Ross 1977a; Leland and Pyle 1977; Bhattacharya 1979). If the managers of the firm possess information that is not held by the market then the market will make inferences from the actions of the firm and in particular, from financial decisions. Changes in its financial structure or its dividend policy will alter investors' perceptions of its risk class and, therefore, its value. While the operator, $E^*(\cdot)$, does not change, the perception of the distribution of the firm's cash flows does. In an effort to maximize their value, firms will take actions, such as taking on high debt to equity ratios, which can be imitated by lesser firms only at a prohibitive cost. This will distinguish them from lesser firms that the uninformed market erroneously puts into the same classes with them. In this fashion, a hierarchy of firm risk classes will emerge, and, in equilibrium, firms will signal their true situations and investors will draw correct inferences from their signals.

All of this has a nice ring to it, but the nagging question that remains is why firms use their financial decisions to accomplish all of this information transfer. Financial changes are cheap, but even cheaper might be guarantees or, for that matter, a system of legislation. These issues remain unresolved, but it is difficult to think that much will be explained by theories that argue that firms take on more debt just to show the world that 'they can do it'. There is a limit to macho-finance.

Taxes

Another line of attack has been to introduce more 'imperfections', especially taxes, into the models. Modigliani and Miller originally had noted that the presence of a corporate tax meant that firms would have an incentive to issue additional debt. Since interest payments on debt are excluded from corporate taxes, substituting debt for equity permits firms to pass returns to investors with a lowered tax cut to the government. At the limit, firms would be all debt if the tax authorities still recognized such debt payments as excludable from taxable corporate income. Presumably, the only brake to this expansion would be the real costs of dealing with the inevitable bankruptcies of high debt firms. This is logically possible, but at the expense of reducing corporate finance to the study of the tradeoff between the tax advantages of debt and the costs of bankruptcy.

Miller (1977) found a more profound brake to this tendency to increase debt. He argued that while the firm could lower its taxes by increasing its debt, the ability of investors to defer or offset capital gains implies that they pay higher taxes on interest income than on the returns from equities. With a rising tax schedule, an equilibrium is possible in which the marginal investor has a tax differential between ordinary income and equity returns that exactly offsets the firm's corporate tax advantage to debt. In such an equilibrium, investors in a higher tax bracket than the marginal investor would purchase only equity (or non-taxed bonds such as municipals for US investors) and those in lower tax brackets would purchase corporate bonds. There would be an equilibrium amount of debt for the corporate sector as a whole, but not for any individual firm (assuming the absence of inframarginal firm tax schedules).

Miller's analysis led to a large literature on the impact of taxes on pricing. Black and Scholes (1974) had made a related argument for the absence of a tax effect on dividends, arguing that stocks with relatively higher yields should not have higher gross returns to compensate investors for the additional tax burden since companies would then just cut their dividends to increase the stock price. Black and Scholes verified their results empirically, but, using a different methodology, Litzenberger and Ramaswamy (1982) found that gross returns were higher for stocks with higher dividends. Whether the supply side or the demand side dominates remains undecided.

Whatever the resolution of this and similar debates, the equilibrium tax argument initiated by Miller has changed much of the analysis of these issues. Miller and Scholes (1978), for example, argue that by employing a number of 'laundering' devices individuals can dramatically cut their taxes. Their conclusion that, in theory, taxes should be much lower than they appear to be in practice, focuses attention on the role played by informational asymmetries and the related

costliness of using techniques such as investing through tax exempt intermediaries.

Agency Models

The emphasis on informational asymmetries has been the cornerstone of an alternative approach to corporate finance, agency theory. Wilson (1968) and Ross (1973) developed agency models in which one party, the agent (e.g. a corporate manager) acts on behalf of another the principal (e.g. stockholders). Jensen and Meckling (1976), building on the agency theory and on Williamson's (1975) transaction cost approach, argue that corporate finance can be understood in terms of the monitoring and bonding costs imposed on stockholders and managers by such relations. The manager qua employee has an incentive to divert firm resources to his own benefit. Jensen and Meckling refer to the loss in value in restraining this incentive as the (equilibrium) agency cost of the relation.

To some extent this conflict can be resolved *ex ante* by the indenture agreements and covenants in financial contracts, but the cost of doing so rises with the monitoring requirements. Myers (1977), for example, has studied the implications for investment policy of the conflict between the stockholders and the bondholders. Stockholders own a call option on the assets of the firm and the value of a call increases with the variance of the asset value. Conversely, such increases will come at the expense of the bondholders. *Ex ante* indenture agreements can limit the ability of management and stockholders to take on additional risk, but the more precise the limits the costlier it is to write, observe and enforce them.

These trade-offs are the intuition and subject matter of the agency approach to corporate finance, but to date it is more a collection of intuitions than a well-articulated theory. The agency approach has pointed in some intriguing directions, but it fares poorly if judged by asking what it is that would be a counter observation or count as evidence against it. To the contrary, no phenomenon seems beyond the reach of 'agency costs' and at times the phrase takes on more of the trappings of an incantation than an analytical tool. The role of asymmetric information in corporate finance and in explaining the managerial and financial forces at work in the firm is self evident, but it remains fertile ground for theory.

Empirical Evidence

The early empirical work examined the relation between the corporate financial structure and other characteristics of the firm. Hamada (1972), for example, studied whether the beta of a firm's equity was related to the beta of the firm's assets as would be predicted by the cost of capital, Eq. 33. There continues to be empirical work on these issues, but the attention of empiricists has shifted to the arena of corporate control.

A boom in merger and acquisition activity in the late 1970s and through to the present time has brought some striking and unexplained empirical regularities. On average, shareholders in firms that are the targets of tender offers gain significantly from such offers while the rewards to bidders are still ambiguous (Jensen and Ruback 1983). For unsuccessful tenders the target firms appear to average an eventual loss and the bidders may, too. These results and the discrepancy between targets and bidders have been the object of close scrutiny.

If firms realize such abnormal gains as targets, and if it reflects the release of information about the value of their underlying assets, then that raises the question of why they were not priced correctly to begin with. On the other hand, if the returns for successful targets reflect synergies rather than simply a revaluation of their assets, why does the bidder get so little? Several game theoretic and bidding models have been built in an attempt to explain these results (see, e.g., Grossman and Hart 1980), but a consensus has yet to emerge. Furthermore, some of the important empirical issues, such as whether bidders actually gain or lose on average remain unresolved.

Conclusion

For corporate finance, like the other major areas of finance, the neoclassical theory is now well established, but, like the other areas, the inadequacy of the neoclassical analysis is pushing

researchers to begin the challenging but promising exploration of theories of asymmetric information. This work holds out the hope of explaining some of the deeper mysteries of finance that have eluded the neoclassical theory, from the embarrassing plethora of anomalies in capital markets to the basic questions of financial structure.

Perhaps the feature that truly distinguishes finance from much of the rest of economics is this constant interplay between theory and empirical analysis. The test of these new approaches will be decided less by reference to their aesthetics and more by their usefulness in explaining financial data. At the height of the subject, these two criteria become one.

See Also

- ▶ Arbitrage
- ▶ Capital Asset Pricing Model
- ▶ Dividend Policy
- ▶ Efficient Markets Hypothesis
- ▶ Finance (New Developments)
- ▶ Options

Bibliography

Admati, A. 1985. A noisy rational expectations equilibrium for multi-asset securities markets. *Econometrica* 53 (3): 629–657.

Bhattacharya, S. 1979. Imperfect information, dividend policy and the 'bird in the hand' fallacy. *Bell Journal of Economics and Management Science* 10 (1): 259–270.

Black, F. 1972. Capital market equilibrium with restricted borrowing. *Journal of Business* 45 (3): 444–455.

Black, F., and M. Scholes. 1972. The valuation of option contracts and a test of market efficiency. *Journal of Finance* 27 (2): 399–417.

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81 (3): 637–654.

Black, F., and M. Scholes. 1974. The effects of dividend yield and dividend policy on common stock prices and returns. *Journal of Financial Economics* 1 (1): 1–22.

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.C. Jensen. New York: Praeger.

Breeden, D.T. 1979. An intertemporal asset pricing model with stochasticconsumption and investment opportunities. *Journal of Financial Economics* 7 (3): 265–296.

Brennan, M.J., and E.S. Schwartz. 1977. The valuation of American put options. *Journal of Finance* 32 (2): 449–462.

Brown, S., and M. Weinstein. 1983. A new approach to testing asset pricing models: The bilinear paradigm. *Journal of Finance* 38 (3): 711–743.

Cass, D., and J. Stiglitz. 1970. The structure of investor preferences and asset returns, and separatability in portfolio selection: A contribution to the pure theory of mutual funds. *Journal of Economic Theory* 2 (2): 122–160.

Chamberlain, G. 1983. Funds, factors and diversification in arbitrage pricing models. *Econometrica* 51 (5): 1305–1323.

Chamberlain, G., and M. Rothschild. 1983. Arbitrage, factor structure and mean-variance analysis on large asset markets. *Econometrica* 51 (5): 1281–1304.

Chen, N., R. Roll, and S.A. Ross. 1986. Economic forces and the stock market. *Journal of Business* 59: 383–403.

Chiras, D.P., and S. Manaster. 1978. The information content of option prices and a testofmarketefficiency. *Journal of Financial Economics* 6 (2–3): 213–234.

Connor, G. 1984. A unified beta pricing theory. *Journal of Economic Theory* 34 (1): 13–31.

Cootner, P., ed. 1964. *The random character of stock market prices*. Cambridge, MA: MIT Press.

Cowles, A. 1933. Can stock market forecasters forecast? *Econometrica* 1: 309–324.

Cox, J.C., and S.A. Ross. 1976a. The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3 (1–2): 145–166.

Cox, J.C., and S.A. Ross. 1976b. A survey of some new results in financial option pricing theory. *Journal of Finance* 31 (2): 383–402.

Cox, J.C., S.A. Ross, and M. Rubinstein. 1979. Option pricing: A simplified approach. *Journal of Financial Economics* 7 (3): 229–263.

Cox, J.C., J. Ingersoll, and S.A. Ross. 1985a. An intertemporal general equilibrium model of asset prices. *Econometrica* 53 (2): 363–384.

Cox, J.C., J. Ingersoll, and S.A. Ross. 1985b. A theory of the term structure of interest rates. *Econometrica* 53 (2): 385–407.

Diamond, D., and R. Verrecchia. 1981. Information aggregation in a noisy rational expectations economy. *Journal of Financial Economics* 9 (3): 221–235.

Dybvig, P. 1983. An explicit bound on deviations from APT pricing in a finite economy. *Journal of Financial Economics* 12 (4): 483–496.

Dybvig, P., and S.A. Ross. 1985. Yes, the APT is testable. *Journal of Finance* 40 (4): 1173–1188.

Fama, E.F. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25 (2): 383–417.

Fama, E.F., and J. MacBeth. 1973. Risk, return and equilibrium: Empirical tests. *Journal of Political Economy* 81 (3): 607–636.

Fama, E.F., L. Fisher, M. Jensen, and R. Roll. 1969. The adjustment of stock prices to new information. *International Economic Review* 10 (1): 1–21.

Flavin, M. 1983. Excess volatility in the financial markets: A reassessment of the empirical evidence. *Journal of Political Economy* 91 (6): 929–956.

French, K. 1980. Stock returns and the weekend effect. *Journal of FinancialEconomics* 8 (1): 55–69.

French, K., and R. Roll 1985. Stock return variances, the arrival of information and the reaction of traders. UCLA Working Paper.

Galai, D. 1977. Tests of market efficiency of the Chicago Board Options Exchange. *Journal of Business* 50 (2): 167–197.

Geske, R. 1979. A note on an analytical valuation formula for unprotected American call options on stocks with known dividends. *Journal of Financial Economics* 7 (4): 375–380.

Geske, R., and R. Roll. 1984. Isolating the observed biases in call option pricing: An alternative variance estimator. UCLA Working Paper, April.

Gibbons, M. 1982. Multivariate tests of financial models: A new approach. *Journal of Financial Economics* 10 (1): 3–27.

Gibbons, M., S.A. Ross, and J. Shanken. 1986. A test of the efficiency of a given portfolio. Stanford University Working Paper No. 853.

Granger, C.W.J., and O. Morgenstern. 1962. Spectral analysis of New York stock market prices. Econometric Research Program, Princeton University, Research Memorandum, September.

Grossman, S.J. 1976. On the efficiency of competitive stock markets where traders have diverse information. *Journal of Finance* 31 (2): 573–585.

Grossman, S.J., and O. Hart. 1980. Takeover bids, the free-rider problem, and the theory of the corporation. *Bell Journal of Economics and Management Science* 11 (1): 42–64.

Grossman, S.J., and J. Stiglitz. 1980. The impossibility of informationally efficient markets. *American Economic Review* 70: 393–408.

Hamada, R.S. 1972. The effect of the firm's capital structure on the systematic risk of common stocks. *Journal of Finance* 27 (2): 435–452.

Hansen, L., and R. Singleton. 1983. Stochastic consumption, risk aversion and the temporal behavior of asset returns. *Journal of Political Economy* 91 (2): 249–265.

Harrison, J.M., and D.M. Kreps. 1979. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20 (3): 381–408.

Hicks, J.R. 1946. *Value and capital*. 2nd ed. London: Oxford University Press.

Huberman, G. 1982. A simple approach to arbitrage pricing theory. *Journal of Economic Theory* 28 (1): 183–191.

Ingersoll, J. 1977. A contingent-claims valuation of convertible securities. *Journal of Financial Economics* 4 (3): 289–322.

Ingersoll, J. 1984. Some results in the theory of arbitrage pricing. *Journal of Finance* 39 (4): 1021–1039.

Jaffé, J. 1974. The effect of regulation changes on insider trading. *Bell Journal of Economics and Management Science* 5 (1): 93–121.

Jensen, M.C., and W.H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3 (4): 305–360.

Jensen, M.C., and R.S. Ruback. 1983. The market for corporate control: The scientific evidence. *Journal of Financial Economics* 11 (1–4): 5–50.

Jobson, J.D., and B. Korkie. 1982. Potential performance and tests of portfolioefficiency. *Journal of Financial Economics* 10 (4): 433–466.

Kandel, S. 1984. On the exclusion of assets from tests of the mean variance efficiency of the market portfolio. *Journal of Finance* 39 (1): 63–73.

Keim, D. 1983. Size-related anomalies and stock return seasonality: Further empirical evidence. *Journal of Financial Economics* 12 (1): 13–32.

Kleidon, A. 1986. Variance bounds tests and stock price valuation models. *Journal of Political Economy* 94 (5): 953–1001.

Leland, H., and D. Pyle. 1977. Informational asymmetries, financial structure andfinancial intermediation. *Journal of Finance* 32 (2): 371–387.

Lintner, J. 1965. The valuation of risk assets and the selection of risky investmentsin stock portfolios and capital budgets. *Review of Economics and Statistics* 47: 13–37.

Litzenberger, R., and K. Ramaswamy. 1982. The effect of dividends on common stock prices: Tax effects or information effects? *Journal of Finance* 37 (2): 429–443.

Lucas, R.E. Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46 (6): 1429–1445.

MacBeth, J., and L. Merville. 1979. An empirical examination of the Black-Scholes call option pricing model. *Journal of Finance* 34 (5): 1173–1186.

Malkiel, B. 1966. *The term structure of interest rates: Expectations and behavior patterns*. Princeton: Princeton University Press.

Markowitz, H.M. 1959. *Portfolio selection: Efficient diversification of investments*. New York: Wiley.

Marsh, T., and R.C. Merton. 1986. Dividend variability and variance bounds tests for the rationality of stock market prices. *American Economic Review* 76 (3): 483–498.

Merton, R.C. 1973a. An intertemporal capital asset pricing model. *Econometrica* 41 (5): 867–887.

Merton, R.C. 1973b. Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4 (1): 141–183.

Milgrom, P., and N. Stokey. 1982. Information, trade and common knowledge. *Journal of Economic Theory* 26 (1): 17–27.

Miller, M.H. 1977. Debt and taxes. *Journal of Finance* 32 (2): 261–275.

Miller, M.H., and M. Scholes. 1978. Dividends and taxes. *Journal of Financial Economics* 6 (4): 333–364.
Modigliani, F., and M.H. Miller. 1958. The cost of capital, corporation finance, and the theory of investment. *American Economic Review* 48: 261–297.
Modigliani, F., and M.H. Miller. 1963. Corporate income taxes and the cost of capital. *American Economic Review* 53: 433–443.
Mossin, J. 1966. Equilibrium in a capital asset market. *Econometrica* 34 (4): 768–783.
Myers, S. 1977. Determinants of corporate borrowing. *Journal of FinancialEconomics* 5 (2): 147–175.
Parkinson, M. 1977. Option pricing: The American put. *Journal of Business* 50 (1): 21–36.
Patell, J.M., and M.A. Wolfson. 1979. Anticipated information releases reflected in call option prices. *Journal of Accounting and Economics* 1 (2): 117–140.
Roll, R. 1970. *The behavior of interest rates: An application of the efficient market model to US treasury bills*. New York: Basic Books.
Roll, R. 1977a. An analytic valuation formula for unprotected American call options on stocks with known dividends. *Journal of Financial Economics* 5 (2): 251–258.
Roll, R. 1977b. A critique of the asset pricing theory's tests. *Journal of Financial Economics* 4 (2): 129–176.
Roll, R. 1978. Ambiguity when performance is measured by the securities market line. *Journal of Finance* 33 (4): 1051–1069.
Roll, R. 1981. A possible explanation of the small firm effect. *Journal of Finance* 36 (4): 879–888.
Roll, R. 1983. The turn-of-the-year effect and small firm premium. *Journal of Portfolio Management* 9 (2): 18–28.
Roll, R. 1984. Orange juice and weather. *American Economic Review* 74 (5): 861–880.
Roll, R., and S.A. Ross. 1980. An empirical investigation of the arbitrage pricing theory. *Journal of Finance* 35 (5): 1073–1103.
Ross, S.A. 1973. The economic theory of agency: The principal's problem. *American Economic Review* 63 (2): 134–139.
Ross, S.A. 1975. Uncertainty and the heterogeneous capital good model. *Review of Economic Studies* 42 (1): 133–146.
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 (3): 341–360.
Ross, S.A. 1976c. Options and efficiency. *Quarterly Journal of Economics* 90 (1): 75–89.
Ross, S.A. 1977a. The determination of financial structure: The incentive signalling approach. *Bell Journal of Economics and Management Science* 8 (1): 23–40.
Ross, S.A. 1977b. The capital asset pricing model (CAPM), short-sale restrictions and related issues. *Journal of Finance* 32 (1): 177–183.
Ross, S.A. 1978a. Mutual fund separation in financial theory – The separating distributions. *Journal of Economic Theory* 17 (2): 254–286.
Ross, S.A. 1978b. A simple approach to the valuation of risky streams. *Journal of Business* 51 (3): 453–475.
Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics and Management Science* 7 (2): 407–425.
Shanken, J. 1982. The arbitrage pricing theory: Is it testable? *Journal of Finance* 37 (5): 1129–1140.
Sharpe, W. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.
Shiller, R. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71 (3): 421–436.
Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25: 65–86.
Williamson, O.E. 1975. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press.
Wilson, R.B. 1968. The theory of syndicates. *Econometrica* 36 (1): 119–132.

Finance (New Developments)

Jiang Wang

Abstract

This survey of some of the developments in finance since the mid-1980s begins with advances in the application of arbitrage pricing, and then expands into areas of general asset pricing under the title 'risk and return'. Limitations in our current understanding of risk as well as more data explorations have led to the 'discovery' of anomalies, which challenge classic notions of market efficiency. We examine recent attempts to expand the neoclassical framework to incorporate market imperfections in asset pricing, which, in their more general forms, take centre stage in advances in corporate finance.

Keywords

Adverse selection; Agency problems; Anchoring; Anomalies; Arbitrage; Arbitrage pricing; Asymmetric information; Bankruptcy; Behavioural finance; Belief perseverance;

Bonds; Capital asset pricing model; Capital structure; Consumption-based capital asset pricing model; Contagion; Control rights; Corporate control; Corporate finance; Corporate governance; Cost of capital; Debt and equity; Derivatives; Efficient market hypothesis; Equilibrium; Equity premium puzzle; Expected utility; Finance; Finance (new developments); Human capital; Incomplete contracts; Interest rates; Intertemporal capital asset pricing model; Inventory cost; Irreversible investment; Jumps; Law of one price; Liquidity; Market frictions; Market microstructure; Mean-variance efficient portfolios; Moral hazard; No arbitrage; Optimal contracts; Option pricing; Overconfidence; Private information; Probability; Reduced form models; Risk; Risk aversion; Risk premium; Serial correlation; Short-sale constraints; Speculative bubbles; State-dependent preferences; Stochastic discount factor; Stochastic volatility models; Structural models; Substitutable securities; Takeovers; Trading costs; Transaction costs; Uncertainty; Venture capital

JEL Classifications
G0

This article attempts to survey some of the developments in finance since the mid-1980s. By then, what we know as neoclassical finance had taken its broad shape and become a foundation for our understanding of the central issues in finance as well as the starting point for further developments. As true for any science, the advances in finance since then are characterized by an interactive process of more rigorous testing and revision of the existing theories, more extensive exploration of the data, old and new, and further expansion of theory beyond the known territories.

Finance is concerned with how the financial market facilitates the allocation of capital or assets. In a well-functioning market, it is prices that guide the allocation. Thus, how to value financial assets or financial securities is a primary focus of finance. Since the value of an asset comes from its future payoff, which extends over time and is uncertain in nature, risk is the key element in asset valuation. Relying on a few basic principles, neoclassical finance has developed a rich set of models and tools for asset pricing, risk analysis and corporate finance.

However, much of the neoclassical finance abstracts away from market imperfections, most notably information asymmetry and market frictions. Such an abstraction draws the boundaries of the neoclassical theory. In particular, the theory's applicability softens when these imperfections are important. Moreover, the theory itself does not provide much guidance in gauging the relevance of imperfections. The omission of imperfections also leaves the neoclassic theory mostly free of institutions. Such an simplification is perhaps most stark in corporate finance, as the very existence and consequently the behaviour of firms are presumably institutional arrangements in response to imperfections, but it is similarly striking in the context of financial market, which consists of a complex and collection of institutions and intermediaries. Naturally, this limits what the neoclassical theory says about the behaviour of institutions as major participants in the market and its implications on market behaviour and capital allocation. It should be emphasized that significance of imperfections in a given context is as much, if not more, of an empirical issue as a theoretical matter. To a large extent, developments beyond the neoclassical theory involve very much the interplay between the research in these two dimensions.

Our discussion of these developments will begin with the advances in the application of arbitrage pricing, arguably the most successful area of modern finance. It then expands into areas of general asset pricing under the title 'risk and return'. Limitations in our current understanding of risk as well as more data explorations have led to the 'discovery' of anomalies, which amounted to new challenges to classic notions of market efficiency. After reviewing some of this empirical evidence, we examine some of the recent attempts to expand the neoclassical framework to incorporate market imperfections. We then turn to advances in corporate finance, in which imperfections, in their more general forms, take centre stage.

This article is intended to provide an update to finance, which remains a timeless piece in capturing the spirit and the essence of neoclassical finance. We will rely on it for a more detailed review of the earlier work as well as their historical context. In order to make the two articles more integrated, we adopt a similar framework, but adjusted to reflect the current landscape. Needless to say, any survey of a broad and fast-growing field such as finance will be partial and incomplete.

Arbitrage Pricing

The first principle of asset pricing is the absence of arbitrage. Here, an arbitrage refers to a set of transactions in the market based on public information that always yields net gains. The intuition for no arbitrage is straightforward: any such an opportunity would be exploited by market participants until it disappears. When transactions of financial securities face little frictions, no arbitrage yields sharp results on the prices of securities which are close substitutes. In particular, securities with same payoffs must have the same price. Classical applications of this principle include arbitrage relations between the prices of default-free bonds with different coupons, spot and forward prices of commodities, and spot and forward exchange rates between currencies and their corresponding interest rates. The path-breaking work of Black and Scholes (1973) and Merton (1973) on option pricing greatly expanded the applications of arbitrage pricing. Vasicek (1977) and Cox et al. (1985b) demonstrated how the arbitrage method can be applied to the pricing of default-free bonds. Merton (1974) and Black and Cox (1976) applied the option pricing technique to value bonds with default risks.

Arbitrage pricing as a general methodology enjoyed unprecedented success in finance and in all of economics. Not only did earlier empirical tests find strong support from the data (see, for example, Black and Scholes 1973), but data converged to theory as deviations disappeared quickly with the theory's dissemination. The explosive applications of the methodology by the financial industry, ranging from new products and markets, new pricing and trading technologies, to new investment and risk management practices, gave the core substance for what is now branded as financial engineering.

For a set of traded securities, let P_t denote their current prices and X_{t+1} their next period payoffs (in vector forms). In its general formulation by Ross (1976) and Harrison and Kreps (1979), the absence of arbitrage is equivalent to the existence of a set of positive state prices φ such that

$$P_t = \sum_{\omega} \varphi_t(\omega) X_{t+1}(\omega) = E_t^* \left[e^{-r_{0,t+1}} X_{t+1} \right] \quad (1)$$

where ω denotes a future state of the economy, $\varphi_t(\omega)$ the state price at t for state ω at $t+1$, $r_{0,t+1}$ denotes the risk-free interest rate from t to $t+1$, and $E_t^*[\cdot]$ denotes the conditional expectation using normalized stated prices as probabilities, which is also referred to as the risk-neutral measure. For a set of securities with payoffs determined by the same set of future states ω, they become substitutes. Their prices will then be related by arbitrage through the corresponding state prices. If we can identify a sufficient number of such securities, then their prices will reveal the corresponding state prices, which then allow us to value other substitutes. Certain securities are natural substitutes, such as an underlying asset and its derivatives, the whole collection of fixed income securities, and a firm's bonds and its equity.

Arbitrage pricing has been widely used in the valuation of these securities.

Equity Options

The basic framework for option valuation was established by the pioneer work of Black and Scholes (1973) and Merton (1973) and the contributions of Cox and Ross (1976a, b) and Cox et al. (1979). The huge body of work that followed has enriched this framework substantially. One focus is to allow for more general price behaviour for the underlying asset. This is in part motivated by deviations in the observed prices from the Black–Scholes formula, which assumes a geometric Brownian diffusion for the price of the underlying asset. For example, from the observed option

prices, the volatility implied from the Black–Scholes formula changes over time and differs for different strike prices, a phenomenon referred to as volatility smiles or smirks. Two natural extensions are to allow stochastic volatility and jumps. Hull and White (1987), Heston (1993) and Stein and Stein (1991) proposed models with time-varying volatility. Extending Merton (1976), Amin (1993), Scott (1997) and Bates (2000) have incorporated jumps into stochastic volatility models. Empirical analysis of the data on both options and underlying equities suggests that both stochastic volatility and jumps are helpful in explain their behaviour (see, for example, Melino and Turnbull 1990; Bates 1996; Bakshi et al. 1997; Pan 2002). In a discrete-time setting, the distinction between stochastic volatility and jumps becomes moot. Rubinstein (1994) has suggested modifications to binomial model of Cox et al. (1979) to accommodate the effects of time-varying price dynamics.

Although most of the recent work on equity options stays within the neoclassical arbitrage pricing framework, it significantly enriches the pricing models to better fit the data. But the fit is never perfect. Is the gap eventually going to be closed with more sophisticated models or revealing something more? We are not totally sure. The arbitrage approaches works when options are truly substitutes of the underlying asset. If so, why do they appear in the first place? Market imperfections may be part of the reason, but the exact nature of this link is far from being well understood.

Default-Free Bonds

Bonds of no default risk are closely related to each other as their prices are all driven by interest rates. From (Eq. 1), the price of a pure discount bond, which has a unit payoff at date T, is given by

$$B_t(T) = E_t^* \left[e^{-\int^{tT} r_{0;s} ds} \right]. \qquad (2)$$

The specification of the interest rate process under the risk-neutral measure will then allow us to price bonds by computing the above conditional expectation. It is well known that bond returns share a small number of common factors (for example, Litterman and Scheinkman 1991). A natural approach is to specify the interest rate as a function of a few state variables. Earlier work chooses the short-term interest rate as the single state variable and assumes tractable dynamics. For example, Vasicek (1977) assumes a Gaussian Markov process for the short rate and Cox et al. (1985b) assume a square-root process (the CIR model). Other candidate models include Brennan and Schwartz (1979), Cox et al. (1980), Courtadon (1982), and more recently Longstaff (1989), Chan et al. (1992), Constantinides (1992) and Ahn et al. (2002). These models aim at tractable solutions to bond prices to capture their basic behaviour.

The empirical evaluation of these models requires the further specification of r_t. under the statistical measure, that is, the true data-generating process. The transformation from the risk-neutral measure to the statistical measure effectively reflects the risk premium. Thus, the test of an arbitrage-based pricing model is really a joint test of the proposed interest rate process and the associated risk premium process.

Analysis by Brown and Dybvig (1986), Chan et al. (1992) and Gibbons and Ramaswamy (1993) readily demonstrates that the parsimony of single factor models also limits their ability to fit the data. Nonparametric tests, such as Ait-Sahalia (1996) and Stanton (1997), further suggest that single-factor diffusion models are unable to capture several important features of interest rate dynamics. Following Langetieg (1980), Cox et al. (1985b) and Schaefer and Schwartz (1984), multifactor extensions became the next pursuit, notably Chen and Scott (1992), Longstaff and Schwartz (1992) and Hull and White (1994).

Despite their added flexibility, multifactor models face two challenges. On the one hand, they quickly become less tractable as more state variables are added. On the other hand, even though a small number of factors, typically three, capture a large percentage of commonality in bond returns, it is far from clear if they are enough in describing bond prices.

The first challenge has largely limited the focus to tractable models. One notable example is the

so-called affine models, in which the short-term interest rate is assumed to be an affine function of a set of state variables. In addition, the vector of state variables follows a diffusion process with its drift and covariance both being its own affine functions. Closed-form solutions can be obtained for bond prices and yields under this specification. Brown and Schaefer (1994), for example, explored an extension of the CIR model under the affine structure. Duffie and Kan (1996) expanded the scope of the affine models and Dai and Singleton (2000) provided an extensive empirical analysis of their pros and cons. They can capture many aspects of the bond price behaviour, but always leaving a few others. This situation is not unique to the affine models as it is shared by other multifactor models outside the affine class (for example, Ahn et al. 2002). Other enrichments have also been considered, such as jumps in interest rates (for example, Johannes 2004; Piazzesi 2005) and regime shifts in interest rate dynamics (for example, Hamilton 1988; Gray 1996), to enhance the descriptive power of the arbitrage-based models.

It might be unrealistic to attempt to describe the rich behaviour of a large cross-section of bond prices by a small number of risk factors following relatively simple dynamics. One possibility is to relax the limit on the dimension of risk factors. Kennedy (1994), Goldstein (2000) and Santa-Clara and Sornette (2001) have explored models with an infinite-dimensional state vector. However, to make these models empirically implementable, restrictive structures need to be imposed.

Aside from specific modelling issues, what we confront is a more general situation. To the extent that default-free bonds are substitutes, we can rely on arbitrage methods in their valuation. While being very close, they are rarely exact substitutes. From this perspective, arbitrage results provide only an approximation. We can take comfort in the empirical success of existing models so far – the bottle is not full, but close to it. But to fill the rest is proven hard. It suggests that old tricks may be inefficient and something new is at play.

Defaultable Bonds

As Black and Scholes (1973) and Merton (1973) observed, a firm's securities, its bonds and equity, all share the same risk, the risk of its asset value. Given the firm's value and its dynamics, we can then value its bonds in the same way as we value equity options. Merton (1974) and Black and Cox (1976) are among the first set of so-called structural models, which rely on the specific risk structure of corporate bonds with respect to its underlying asset to value them using arbitrage methods. More comprehensive models were further developed to incorporate richer risk structures embedded in corporate bonds. For example, Longstaff and Schwartz (1995) and Saa-Requejo and Santa-Clara (1999) allow interest rate risks, which can affect when a bond defaults and its value then. Richer debt structure, the cost of default and shareholders optimal financing and default choices are also considered by Leland (1994), Anderson and Sundaresan (1996), and Leland and Toft (1996), among others.

Structural models, coupled with specific description of the firm value dynamics, also impose certain behaviour on the event of default. The desire to have more flexibility in modelling the default event has led to the development of so-call reduced form models, for example, Jarrow and Turnbull (1995), Lando (1998) and Duffie and Singleton (1999). These models start by modelling the default process and recovery rate under the risk-neutral measure. From (Eq. 1), the price of a defaultable bond with zero coupon and unit par value is given by

$$P_t(T) = E_t^*\left[e^{-\int_{tT} r_{0;s}ds} 1_{\{\tau>T\}}\right] + E_t^*\left[e^{-\int_{tT} r_{0;s}ds} z_\tau 1_{\{\tau>T\}}\right] \quad (3)$$

where τ denotes the default time, Z_τ the recovery rate in the case of default and $1\{\cdot\}$ is an indicator function. The implementation and evaluation of the reduced form models requires additional information on default events under both the risk-

neutral and the statistical measures. Such information is scarce as default is relatively infrequent. However, these models become more applicable for credit derivatives when more securities related to the same default events became available.

Other Derivatives

The application of arbitrage pricing finds its most fertile ground in valuing derivatives, which, together with the underlying asset, are close substitutes. Its success has fuelled the big bang of derivatives since the 1970s, which provided new areas for more applications of the theory. Commodity and financial futures, interest rate derivatives such as swaps, caps and swaptions, credit derivatives such as credit default swaps (CDS) and collateralized debt obligations (CDO) are major examples.

While no arbitrage as a theoretical principle is quite general, its usefulness in asset valuation remains a practical matter. It applies to financial securities which are close substitutes, when market frictions are negligible. Its success in asset pricing as opposed to in pricing physical goods very much reflects the fact that financial securities are easy to trade and replicate and they are close in nature in the sense that their value all arises from their financial payoffs. Nonetheless, market frictions do exist. Moreover, except in certain instances, securities of interest are often not exact substitutes. The theory is much less definitive about how it extends to these situations.

Its limitations are evident even in areas where it was successful, such as equity options, bonds and other derivatives. Nagging deviations from arbitrage-based models, though arguably small, persist. The need to address these deviations tends to push for more complex models with more risk factors. However, a blind pursuit in this process might be losing not only the empirical ground, as data becomes less sufficient in supporting the models, but also the theoretical ground. The deviations may well reflect the influence of market frictions or other factors, which are beyond the 'limits' of arbitrage arguments.

Another limitation of the arbitrage approach, perhaps a more fundamental one, is that it takes the risks and their risk premia, that is, the relevant states and the corresponding state prices, as given in establishing price relations among substitutable securities. From a broader perspective, it is important to understand the economic underpinnings of different risks and their pricing implications. Such an understanding may provide the basis to further improve arbitrage pricing models. More importantly, it will allow us to value assets more broadly.

Risk and Return

The broader principle in asset pricing is market equilibrium, which requires that security prices must equate supply and demand. This approach is general since it focuses on how security prices are determined by economic fundamentals, which drive supply and demand. Its application, however, faces several challenges. We need to first specify what constitute the fundamentals. We also need to determine how these fundamentals influence the supply and demand for securities and ultimately their prices. Additional structure on the fundamentals is also needed before we can arrive at useful results.

A key fundamental is the risk characteristics of asset payoffs. We start with the pricing Eq. (1). Suppose that given the state of the economy at t, the conditional probability for state ω at $t + 1$ is $p_t(\omega)$. We can then rewrite (Eq. 1) as

$$\begin{aligned} P_t &= \sum_{\omega} P_t(\omega) \frac{\varphi_t(\omega)}{P_t(\omega)} X_{t+1} \\ &= E_t[m_{t+1} X_{t+1}], \end{aligned} \tag{4}$$

where E_t is the expectation under the actual probability measure given the state at t and $m_{t+1} = \varphi_t(\omega)/p_t(\omega)$ is referred to as state price density or the stochastic discount factor (SDF). Realizing that $X_{t+1}/P_t - 1 + r_{t+1}$ where R_{t+1} is the security's return from t to $t + 1$, we can also re-express (Eq. 4) as

$$1 = E_t[m_{t+1}(1 + r_{t+1})]. \tag{5}$$

A slight variation of (Eq. 5) gives a commonly used expression for the pricing Eq. (4):

$$E_t[r_{t+1}] - r_{0,t+1} = -\mathrm{Cov}_t\left[m_{t+1} = E_t\left[m_{t+1}\right], r_{t+1} - r_{0;t+1}\right], \tag{6}$$

where Cov_t denotes the conditional covariance. Eq. (6) suggests that we can decompose an asset's risk into two components:

$$r_{t+1} - r_{0,t+1} - \left(E_t\left[r_{t+1}\right] - r_{0,t+1} + = a_t(1 - m_{t+1})/E_t\left[m_{t+1}\right] + u_{t+1}. \right. \tag{7}$$

The first component, which is correlated with the SDF, will influence the asset's expected return or its price. The second component, which is uncorrelated, does not. Such a decomposition clearly reveals that risks come in two types, 'priced' and 'not priced'. The amount of an asset's priced risk is measured by a_t, its covariance with the SDF, which determines its risk premium:

$$E_t[r_{t1}] - r_{0,t+1} = a_t\sigma^2_{mt}, \tag{8}$$

where σ^2_{mt} is the conditional variance of the SDF. Knowing the SDF will allow us to specify the priced risk and its premium. Thus, the goal for a general asset pricing theory is then to determine the SDF.

Two differently approaches have been followed in developing models for the SDF. The first approach is to start from the primitives of the economy such as asset payoffs and investors' preferences, derive the asset demand (and supply) and finally arrive at the equilibrium SDF. The second approach is to start from the equilibrium, rely on certain properties of the equilibrium to arrive at the SDF. The first approach has more micro-texture to it and often leads to sharper specifications of the SDF, while the second approach allows more flexibility but with less microeconomic basis.

Factor Models of the SDF

The well-known capital asset pricing model (CAPM) follows the first approach. Under the mean-variance framework of Markowitz (1952) and Tobin (1958), all investors will hold mean-variance efficient portfolios. Such a commonality in investors' asset demand implies that, in equilibrium, the market portfolio, which represents the total supply of assets, must be a mean-variance efficient portfolio. This insight has led Sharpe (1964) and Lintner (1965) to identify the return on the market portfolio $r_{M,t+1}$ to be a proxy for the SDF:

$$1 - m_{t+1}/E_t[m_{t+1}] = b_{M,t}\left(r_{M,t+1} - E_t\left[r_{M,t+1}\right]\right),$$

which immediately leads to the CAPM:

$$E_t\left[r_{i,t+1}\right] - r_{0,t+1} = \beta_{iM,t}\left(E_t\left[r_{M,t+1}\right] - r_{0,t+1}\right), \tag{9}$$

where $\beta_{iM,t}$ is the market beta for asset i. Ross (1976a, b) started directly from the risk structure of asset payoffs. By proposing a linear factor model for asset returns and requiring the absence of limiting arbitrages as an equilibrium condition, he obtained a factor representation for the SDF:

$$1 - m_{t+1}/E_t\left[m_{t+1}\right] = \sum_{k=1}^{K} b_{k,t} f_{k,t+1}, \tag{10}$$

which leads to the arbitrage pricing theory (APT):

$$E_t\left[r_{i,t+1}\right] - r_{0,t+1} = \sum_{k=1}^{K} \beta_{ik,t}\,\lambda_{k,t}\,, \tag{11}$$

where $\beta_{ik,t} = \sigma_{ik,t}/\sigma^2_{k,t}$ is the 'beta' of asset i on risk factor k (that is, the regression coefficient of its return on $f_{k,t+1}$), $\sigma^2{}_{k,t}$ the conditional variance of factor k, $\sigma_{ik,t}$, is its conditional covariance with the return of asset i, and $\lambda_{k,t} = b_{k,t}\sigma^2{}_{k,t}$ its risk premium. Thus, both the CAPM and the APT can be viewed as the case when the SDF has a linear factor structure. The key distinction is that the CAPM identifies the market return as the SDF while the APT allows the SDF to be spanned by multiple factors.

Earlier empirical tests found some supporting evidence for the CAPM (see, for example, Black et al. 1972; Fama and MacBeth 1973). Due to the

F

noise in estimating expected returns and the difficulty to actually identify the market portfolio, questions regarding the strength of the support as well as the nature of these tests have left a strong need for more tests (see, for example, Roll 1977). Further empirical exploration also reveals evidence that is at odds with the CAPM, at least on the face of it. Banz (1981) discovered the size effect that small stocks (measured by market capitalization) yield higher average returns than large stocks, after controlling for what the CAPM predicts. Basu (1983) reported the value effect that stocks with book values higher than their market values – the book-to-market ratio – yield higher average returns than stocks with lower ratios. In a comprehensive empirical analysis, Fama and French (1992) synthesized this evidence, demonstrating the weak explanatory power of the CAMP for the cross-section of stock returns as well as certain patterns they display.

The APT allows a richer structure than the CAPM. However, the fact that the theory itself does not identify the factors poses challenges to its empirical testing (see, for example, Shanken 1982; Dybvig and Ross 1985). Other means need to be used, theoretical or statistical, to identify the factors. Earlier empirical tests along this route find some supportive evidence (for example, Chen et al. 1986) but leaves more to be desired. Relying on statistical analysis, Connor and Korajczyk (1988) use principal components and Lehmann and Modest (1988) use factor analysis to empirically identify the factors. The evidence in support of the APT is, however, mixed. Exposures to the empirically identified factor risks explain only part of the cross-sectional variation in average returns. Based on the observed average returns, Fama and French (1993) propose to use firm characteristics such as size and book-to-market ratio to form portfolios, whose returns are then used to identify risk factors in addition to the market. They show that the size factor (the difference in returns from small and large stocks), the value factor (the difference in returns from high and low book-to-market stocks), plus a broad market index can explain most of the cross-sectional variation in returns from portfolios sorted on by their loadings.

What to take away from the empirical results remains a matter of active discussion. The appeal of the CAPM has led to continuous efforts to reconcile it with the empirical evidence. All empirical implementations of the CAPM use a market index as a proxy for the market portfolio, which leaves the possibility of misidentifications. The lack of significant gains from improving market proxies based on traded assets, as Stambaugh (1982) and Shanken (1987) have shown, has led to the inclusion of non-traded assets. Using labour income growth to measure the return on human capital, Jagannathan and Wang (1996) showed that including human capital in the market proxy may help to increase the explanatory power of the CAPM.

In general the CAPM gives a conditional relation between risk, as measured by an asset's conditional beta, and its conditional risk premium, as (Eq. 9) clearly indicates. However, earlier tests are mostly unconditional, looking at the relation between assets' unconditional beta and their unconditional premia. By allowing time-varying betas and the market premium, more tests are directed at the conditional version of the CAPM, notably Harvey (1989), Shanken (1990), Jagannathan and Wang (1996), and, more recently, Wang (2003) and Petkova and Zhang (2005). How far the added flexibility from the conditional variables and their impact can lead us remains to be seen, as Lewellen and Nagel (2006) demonstrate. More fundamental questions exist about the test of conditional CAPM. For example, what are the appropriate conditioning variables implied by the model? Without knowing this, how do we distinguish the impact of conditioning variables from additional risk factors?

The added flexibility in the empirical multifactor models like that of Fama and French (1992, 1993) also leave plenty of room for alternative interpretations. It is open to potential dangers of data snooping, as Lo and MacKinlay (1990) and Kothari et al. (1995) point out. It can also be a result of misidentification of the true risk factor even when the conditional CAPM holds (see, for example, Berk et al. 1999; Gomes et al. 2003).

The Intertemporal CAPM

Merton (1973) developed an intertemporal version of the CAPM (ICAPM), which shows that time-varying market conditions give rise to dynamic risk factors in addition to the risk of the market portfolio. In particular, if we let the first factor in (10) be the return on the market portfolio, the other factors will represent the state variables driving the market conditions. The ICAPM contains the conditional CAPM as the special case when the dynamic risks carry no premium. In this sense, any test of the conditional CAPM can be viewed as a test of the ICAPM with additional restrictions on risk premia. However, in the ICAPM, the dynamic risk factors are taken as given, not derived from the theory. The pricing relation, in the form of (11), comes as an equilibrium condition under a given form of price dynamics rather than an equilibrium outcome in terms of economic primitives. In this regard, the ICAPM has more in common with APT than with the classical CAPM.

The ICAPM provides a useful framework for analysing risk and return in an intertemporal setting. Its empirical implementation has been limited until recently. Tests of the APT, which also allows multiple risk factors, were often interpreted as tests of the ICAPM, since both models allow plenty of flexibility in identifying these factors. However, such a view leaves out the additional implications from the ICAPM on the intertemporal properties of the dynamic risks. Lo and Wang (2006) have developed a version of the ICAPM in which the time-varying market risk premium captures the dynamic risk. Using the cross-sectional data on trading volume, they empirically identify the dynamic risk factor and test its power in explaining the time series and the cross section of asset returns. Ang et al. (2006) and Adrian and Rosenberg (2006) also test a version of the ICAPM in which the market volatility is a dynamic risk factor.

By identifying risk factors other than the market proxy as dynamics risks, the ICAPM gives more guidance on the properties of these additional risk factors, in particular, their correlation structure with the underlying state variables. These properties may help the empirical construction of these factors and their tests. More work along this direction is called for.

Consumption-Based CAPM

The APT and the ICAPM focus on the statistical properties of risks, in particular, their correlation structure, both in cross-section and time series. But from a pricing perspective, their economic properties are particularly important. Clearly, investors' attitude towards different risks also matter. For an investor, an asset is riskier if its return co-varies more strongly with his future marginal utility. Using the same setting as Merton's ICAPM, Breeden (1979) showed that for pricing purposes, all risks can be collapsed into one, measured by assets' covariance with aggregate consumption (see also Rubinstein 1976). As a result, the beta of an asset with respect to aggregate consumption, the consumption-beta, determines its risk premium. This is the so-called consumption-based CAPM (CCAPM).

Let us start with the principle of market equilibrium, namely, asset prices must equate demand and supply. Since demand and supply for assets are determined by the fundamentals through market participants' optimizing behaviour, so will be their equilibrium prices. Consider a representative investor in the market who has a time-separable, expected utility function $u(c_t) + \rho u(c_{t+1})$. The optimality of his asset holdings requires that $u'(c_t)P_t = \mathrm{E}_t[\rho u'(c_{t+1})X_{t+1}]$ *or*

$$1 = \mathrm{E}_t\left[\frac{\rho u'(c_{t+1})}{u'(c_t)}(1 + r_{t+1})\right]. \qquad (12)$$

Comparing (Eq. 12) with (Eq. 5), it is apparent that market equilibrium imposes additional restrictions on asset prices. In particular, it relates the stochastic discount factor to the marginal utilities of the representative investor: $m_{t+1} = \rho u'(c_{t+1})/u'(c_t)$.

The simple structure of the CCAPM and its economic appeal has generated a lot of interest in its empirical implementation, lead by

Breeden (1980), Grossman and Shiller (1981) and Hansen and Singleton (1983). The focus has been on two fronts, the behaviour of aggregate prices and the cross section of asset returns. The procedure typically involves an estimation of the consumption process and certain specification of the marginal utility function for the representative investor. Combining the two gives an estimate for the SDF, which can then be used to test its pricing implications.

From (Eqs. 7 and 8), Hansen and Jagannathan (1991) established the following condition on the SDF:

$$\sigma_t\left(m_{t+1}\right)/\mathrm{E}_t[m_{t+1}] \geq \frac{\mathrm{E}_t\left[r_{i,t+1}\right] - r_{0,t+1}}{\sigma_t\left(r_{i,t+1} - r_{0,t+1}\right)}, \quad (13)$$

where i denotes any traded asset. The right-hand side is asset i's Sharpe ratio. Thus, the maximum Sharpe ratio of traded assets gives a lower bound for the volatility of the SDF. In the CCAPM, the volatility of the SDF comes from the volatility of aggregate consumption. However, the aggregate consumption data seems very 'dull', exhibiting close to i.i.d. growth with very low volatility. This poses certain challenges to the simple forms of the CCAPM. In order for the SDF to have the desired properties, the marginal utility function has to do all the work. Using a time-separable utility function with constant relative risk aversion, Mehra and Prescott (1985) show that the implied risk aversion has to be very high to yield a volatility of the SDF that exceeds the bounds posted by the Sharpe ratio of the market index. Such a high implied risk aversion seems inconsistent with other evidence on individual risk preferences, leaving us with what is referred as the 'equity premium puzzle'. Applying the model in such a manner quickly led to many more 'puzzles'. Real interest rates have been low, at least in developed markets, which is inconsistent with the observed consumption growth and a high risk aversion (for example, Weil 1989). Also, the low variability in the SDF and the low observed variability in aggregate dividend growth are at odds with the high observed volatility in aggregate asset prices or asset returns (for example, LeRoy and Porter 1981; Shiller 1981; Campbell and Shiller 1988).

Many efforts have been made to reconcile the volatility on the SDF implied by asset returns and the consumption data, mostly along three directions. The first is to improve on the measure of consumption risk. For example, Rietz (1988) suggests that the small probability events like severe drops in consumption may be important risks not fully captured by the data. Bansal and Yaron (2004) propose to use the variability in long-horizon consumption growth as a measure of risk. These explorations are useful, but also stretches the boundaries of the data (for example, a finite sample period will limit the length of the horizon).

The second direction is to allow for more general preferences. The simple form of utility function assumed for the representative investor in early studies is probably too restrictive. For example, even if the preferences of individual investors are restricted to simple forms – such as those exhibiting constant relative risk aversion – a certain degree of heterogeneity will lead to a more complex preference at the aggregate level, which depends on the relative importance of each investor (see, for example, Dumas 1989; Wang 1996; Chan and Kogan 2002). As more flexibility is needed in fitting the data, different forms of state-dependent preferences were considered, notably Sundaresan (1989), Abel (1990), Constantinides (1990), Epstein and Zin (1991), and Campbell and Cochrane (1999), allowing factors like habit, aggregate consumption, and the timing of risk resolution to influence behaviour. The flexibility this approach enjoys also comes at certain costs. On the theory side, the aggregation properties of simple preferences are lost, which leads to questions about the link between what is assumed for the representative agent and the micro justifications used to motivate the preference structure. On the empirical side, there is a lack of discipline in identifying the true preference structure.

The third direction is to allow for certain forms of market imperfections. This approach opens up many possibilities but goes beyond the neoclassical setting. We will return to market imperfections in the next section.

The CCAPM has also been applied to explain the cross section of asset returns, as the other

pricing models. From $m_{t+1} = \rho u'(c_{t+1})/u'(c_t)$, m_{t+1} can be approximated by

$$1 - m_{t+1}/E_t[m_{t+1}] \approx b_t\,(\Delta c_{t+1} - E_t[\Delta c_{t+1}]), \tag{14}$$

where Δc_{t+1} denotes the aggregate consumption growth. The cross section of asset returns are then given by a formula similar to (Eq. 9) with the exception that the beta is now replaced by the consumption beta, the beta of assets' payoff with respect to aggregate consumption growth. Lettau and Ludvigson (2001) have implemented the CCAPM in this form and find that it can explain the cross-sectional pattern in portfolio returns as presented by Fama and French (1992, 1993). Bansal et al. (2005) also find encouraging signs in using assets' betas of their long-run dividends with respect to the long-horizon consumption growth to explain the cross section of their returns.

The appeal of the consumption-based CAPM mainly comes from its simple economic structure. However, its validity relies on strong assumptions about the behaviour of market participants and the structure of the market. It is unclear how much the behaviour of major market participants, such as institutional investors and delegated money managers, is related to consumption. It is also unclear whether the existing market structure allows the kind of efficiency in risk allocation and the proper aggregation needed for the CCAPM. Market imperfections will cause deviations from these assumptions, which may well contribute to the challenges in fitting the model to data. On the empirical side, it is worth pointing out that estimates on risk premium and consumption risk are fairly rough.

Market Efficiency and Anomalies

The neoclassical theory of asset pricing relies on two simplifications, namely, frictions are negligible in financial markets and information is reasonably homogenous among market participants. While the second simplification is less relevant for arbitrage pricing, both are needed for equilibrium-based pricing models. The idea that market participants have similar information regarding future asset payoffs is closely related to the notion of financial markets being informationally efficient, a hallmark of neoclassical finance. The efficient market hypothesis (EMH) postulates that market prices fully reflect all the relevant information available in the market (see, for example, Fama 1970). The intuition behind the hypothesis is simple, very much in the spirit of no arbitrage. Any available information that is not properly reflected in prices will be taken advantage of by profit-seeking market participants, sometimes referred to as arbitrageurs, until prices fully adjust for it.

The exact formulation of the hypothesis, however, involves important subtleties, including the precise definition of relevant information and its reflection in prices. Perhaps the simplest formulation of EMH is to assume the presence of arbitrageurs with unlimited risk tolerance and access to capital (see, for example, Samuelson 1965). This then implies that current prices are unbiased forecasts of future prices (adjusted for time value). Event studies found broad support for prices reacting quickly and quite accurately on average to public news. Extensive tests of predictability found the evidence to be largely consistent. Nonetheless, deviations exist. A natural way to account for the deviations is to relax the condition of risk-neutrality and properly account for risks. After all, (imperfect) predictability does not imply arbitrage, as apparent in Lucas (1978). But such an approach immediately leads us to the choice of a particular method of risk adjustment or an asset pricing model. Tests of EMH then becomes tests of the asset pricing model used, which complicates the matter substantially.

Perhaps as a reaction to the incredible success of the EMH, the initial empirical support was followed by the recording of an increasing number of exceptions, which are also referred to as anomalies. The earlier set of results is about the predictability on equity returns. DeBondt and Thaler (1985) studied long-horizon returns over three to five years. They examined two portfolios, a winner portfolio and a loser portfolio, which consist of stocks with higher or lower than market adjusted returns, respectively, and found that the

winner portfolio yields lower returns in the following years and the loser portfolio yields higher returns. Fama and French (1988) and Poterba and Summers (1988) also found negative serial correlation in long-horizon market index returns. While it is harder to make inferences from long-horizon returns as the sample size becomes relatively small, Lo and MacKinlay (1988) and others have documented positive serial correlation in market index returns over weekly and monthly horizons. This evidence suggests that stock prices need not follow random walks as the weak form of the EMH claims. It was also documented that returns of large stocks can predict returns of small stocks on weekly and monthly basis, the so-called lead–lag phenomenon (see, for example, Lo and MacKinlay 1990). Bernard and Thomas (1989), extending the earlier work by Ball and Brown (1968) and Jones and Litzenberger (1970), have presented convincing evidence of the underreaction of stock prices to earnings announcements, which is later called the 'post earnings announcement drift'. Many studies, such as Campbell and Shiller (1988) and Fama and French (1988), have also suggested that financial ratios such as dividend yield can predict aggregate market returns. These results are certainly at odds with the semi-strong form of the EMH, which requires no predictability of asset returns using public information. Jegadeesh and Titman (1993) take the winner–loser comparison to individual stocks over shorter horizons. By sorting stocks on returns over past one or two quarters, they show that winner portfolio continues to yield higher returns while loser portfolio yields lower returns, a phenomenon called 'momentum'. If we take long positions in winners and offsetting short positions in losers, the average return can be substantial. This is particularly intriguing as we expect diversification and cancellation to greatly limit the net risk exposure of this strategy.

The search for predictability in stock returns has also gained momentum of its own. Different variables were found to have predictive power for equity returns such as trading volume (for example, Gervais et al. 2001), short interest (for example, Jones and Lamont 2002), share repurchases (for example, Ikenberry et al. 1995), dispersion in analysts forecasts (for example, Diether et al. 2002), and transactions of institutional investors (for example, Chan and Lakonishok 1995). The list goes on and may continue to grow. However, several caveats always accompany these findings. First, their significance, both statistical and economical, is quite moderate. Second, their persistence over time needs further testing. Third, more work is desired to distinguish them from potential spurious findings due to data mining.

A corollary of the EMH is that news on future payoffs or the SDF move prices. Roll (1988) showed that *ex post* public news can only explain a fraction of price movements of individual stocks over daily to monthly horizons (see also Roll 1984a). This result parallels what is observed at the aggregate level by LeRoy and Porter (1981) and Shiller (1981), that is, aggregate market indices exhibit a volatility much higher than the volatility in aggregate dividends in the data. But it is more striking as risk considerations seem to be less important over short horizons. The inability to explain price movements even *ex post* has been viewed as a serious challenge to neoclassical theory, in particular the EMH. One possible explanation is that much of the price movement is driven by private news which is not captured by the information set used in the empirical studies. Another is the influence of time-varying risk which may contribute to movements in the discount factor.

To avoid the complication of risk, some studies have focused on more direct tests on the foundation of the EMH, the principle of no arbitrage. A longtime puzzle along this line is the significant discount on closed-end funds from their net asset values (see, for example, Malkiel 1977; Lee et al. 1991). In violation of the law of one price, anomalies of this nature document price differences between two seemingly identical assets. Other well-known examples include the price differences between the on-the-run and the off-the-run Treasury bonds with close to identical payoffs and shares of the same company with the same dividend streams but traded on different exchanges. A pair trade, to buy the security with lower price and sell the price with the higher price, which requires no private information and substantial

capital but yields sure profits, seems to present an arbitrage opportunity. Obviously, the persistent existence of these price anomalies suggests that there is more to what meets the eye. For example, Ross (2002) has shown that management fees contribute significantly to the close-end discounts.

How to interpret the empirical anomalies, assuming their presence, requires further assessment. In the whole, as the name suggests, anomalies do not overweight the vast positive evidence in support of the EMH. Additional factors also need to be included in the consideration. First, predictable patterns in returns or deviations from the law of one price documented in the data are not equivalent to actual profitable opportunities in the market. Frictions in the market need to be taken into account (see, for example, Tuckman and Vila 1992). Second, strategies attempted at taking advantage of these anomalies always involve certain risks in the presence of frictions. The dynamic nature of these risks make them harder to assess (see, for example, Merton 1981; Dybvig and Ross 1985).

Nonetheless, these anomalies, together with deviations in asset returns from neoclassical asset pricing models, do pose a challenge to our understanding of how the market works and how asset prices are determined. It is clear that the notion of market efficiency need to be examined in an equilibrium asset pricing framework, which allows for information asymmetry (and possibly market frictions). Grossman (1976) and Grossman and Stiglitz (1980) demonstrated that such a framework is much richer than the simple form of market efficiency implies, for both the behaviour of asset prices and the importance of information asymmetry in determining it. However, many of the implications of this framework needed to be fleshed out, which became a fertile ground for recent work.

Market Imperfections

Limitations of the neoclassic theory have led to efforts to incorporate imperfections into our analysis, in particular, frictions and asymmetric information. Imperfections influences how the market operates at two levels. At a superficial level, imperfections directly affect why and how investors trade in the market, which ultimately determine asset prices. At a more fundamental level, imperfections also determine the institutional structure of the market itself as well as the economic characteristics of major market participants, both of which also contribute to the actual imperfections observed in the market. Although efforts have been made in analysing imperfections at both levels, more of them focused on the former.

F

Market Frictions

Despite the relative ease of transactions in the financial market, frictions exist. They range from simple trading costs such as commissions and bid–ask spreads to price impact, costs and restrictions on short sales, constraints on borrowing to simple inability to trade-certain claims or contracts. They also include the costs of setting up trading operations, gathering and processing information, maintaining market presence and the costs of introducing a new security, creating and maintaining a market and providing liquidity for it. Since frictions hinder the efficient allocation of capital in the market, their impact is closely related to the notion of liquidity or illiquidity.

Factoring in market frictions sheds new light on the empirical anomalies. Many of them do not provide profitable trading opportunities when trading costs are included. For example, Krishnamurthy (2002) finds that costs in financing the arbitrage between the on-the-run and off-the-run bonds are substantial and outweigh the potential gains. Lesmond et al. (2004), among others, show that momentum in individual stocks is not profitable after adjusting for trading costs. These results are comforting for the EMH and in many ways not surprising. But they do not settle all the questions. In particular, why are these patterns there in the first place, and how do they fit into the overall asset pricing framework?

The general impact of market frictions on asset prices is hard to analyse as they make the behaviour of market participants, the interaction among them and the equilibrium outcome very complex. Recent studies have mainly focused on how specific frictions such as transactions costs, short-sale

constraints and borrowing constraints may influence three aspects of asset prices, the overall level, the cross section and dynamics.

Relying on partial equilibrium arguments, earlier work has examined how transactions costs may influence the level of asset prices. For example, Constantinides (1986) considered the equivalent price adjustment to offset the welfare loss from proportional transactions costs and found that its magnitude is of higher order of the cost and quantitatively insignificant. Amihud and Mendelson (1986) calibrated the present value of implied transactions costs using observed stock trading volume and showed it to be substantial. What is not fully incorporated in the partial equilibrium analysis is how costs affect the actual equilibrium. Vayanos (1998) considered a general equilibrium model in which investors trade for life-cycle reasons and reached the same conclusion as Constantinides. This is not surprising since life-cycle considerations generate little trading and consequently transactions costs have a limited effect. When high levels of trading are needed, as observed in the market, the situation can be different. Unable to trade frequently, investors have to bear additional risks they could otherwise unload in the market, which can significantly alter their behaviour. Allowing high frequency trading needs compatible with observed volume, Lo et al. (2004) show that moderate fixed transactions costs can have a significant impact on investors' asset demand and the resulting equilibrium prices.

In the context of consumption-based CAPM, incorporating market frictions can potentially help to reconcile a high-risk premium with a smooth consumption path (see, for example, He and Modest 1995; Luttmer 1999). The mechanism is quite straightforward. In the presence of frictions, the equality in (Eq. 12) is in general replaced by inequalities. For example, with proportional transaction costs κ and no short sales and borrowing, (Eq. 12) becomes

$$\mathrm{E}_t\left[\rho\frac{u'_{t+1}(c_{t+1})}{u'_t(c_t)}(1+r_{t+1})\right]\leq\frac{1+\kappa}{1-\kappa}, \qquad (15)$$

which loosens the link between prices and marginal utilities. However, using an equilibrium model calibrated to the trading needs to households' heterogeneous labour income risks, Heaton and Lucas (1996) found that transactions costs have a limited effect on the equilibrium risk premium because trading is very moderate in consumption-based models, but trading restrictions such as short-sale and borrowing constraints can potentially have larger effects (see also Constantinides and Duffie 1996; Constantinides et al. 2002; Brav et al. 2002).

How market frictions affect the cross section of asset returns is a challenging issue. Merton (1987) considered an extension of the CAPM in which investors invest only in a subset of assets due to the information cost of learning about them. He showed that the segmentation of the market leads to modifications to the CAPM which exhibit a complex structure, depending on investor preferences, endowments and the nature of the segmentation. Here, more empirical guidance can be helpful. Using various measures of liquidity for individual stocks, Brennan and Subrahmanyam (1996) have documented an empirical link between liquidity and average returns. Liquidity of individual assets seems to exhibit commonalities (see Chordia et al. 2000). This suggests the possibility that liquidity may contain factor risks. Assuming the CAPM to hold net of costs, Acharya and Pedersen (2005) allowed the effective costs in asset trading to be correlated with market returns to help explain the deviations in observed, pre-cost returns from the CAPM. Pastor and Stambaugh (2003) directly include the market average of a liquidity measure proposed by Campbell et al. 1993 as an additional risk factor in the SDF and find that it can enhance the explanatory power of multifactor models. Much is needed for the theoretical basis of the connections between frictions and the cross section of asset returns.

From a theoretical point of view, market frictions can contribute to the dynamic properties of asset prices. When flow of capital is costly, for example, the risk tolerance of marginal investors may increase and become dependent on market conditions, which can lead to predictable asset returns and more volatile prices. For example,

Grossman and Vila (1992) showed that borrowing constraints can force risk-neutral investors to behave in a risk-averse manner. Grossman and Miller (1988) emphasize the imperfect mobility of capital by imposing costs on maintaining market presence and demonstrate that these costs lead to limited risk tolerance in the market and mean reversion in returns when trades are not perfectly synchronized. Using return and volume to infer order imbalances in the market, Campbell et al. 1993 find that they indeed generate return reversals. Pagano (1989) and Allen and Gale (1994) also argue that costly participation in the market can exacerbate price volatility driven by demand shifts over long horizons. Huang and Wang (2006a, b) further point out that low capital mobility in the form of costly participation in the market can lead to endogenous order imbalances. Moreover, the endogenous order imbalances tend to be asymmetric and large when they occur, leading to market crashes, fat-tails in asset returns and return reversals. There is now growing empirical evidence suggesting the low mobility of capital (for example, Coval and Stafford 2007; Mitchell et al. 2007).

Constraints can also influence asset price dynamics. E. Miller (1977a, b) and Harrison and Kreps (1979) have shown that short-sale constraints can inflate asset prices as they prevent short positions and thus can increase asset demand. Scheinkman and Xiong (2003) further demonstrated that short-sale constraints can lead to bubbles and high volatility in prices. Basak (1995) and Grossman and Zhou (1996) show that wealth constraints on market participants can lead to positive correlation between their risk tolerance and price movements. Such a correlation can contribute to higher and more persistent price volatility and mean reversion in returns.

Frictions have also been considered in explaining many other pricing anomalies. For example, Duffie (1996) and Vayanos and Weill (2006) examine how costs in trading from searching in the market can help to explain the price premium and the specialness (that is, high borrowing cost) of on-the-run Treasury bonds. Chen et al. (2002) attempt to associate individual stock return momentum with short-sale constraints. Short-sale constraints can also help to explain empirical findings relating short interests, volume, and dispersion of analysts' forecasts to future returns. Kyle and Xiong (2001) suggest that capital constraints can be the cause of market contagion, which refers to negative co-movements across markets in the absence of negative news affecting both markets.

Although most of the literature has focused on how frictions in the market affect asset demand and consequently prices, some work has also examined the asset supply side, in particular how frictions in firms' real investments may affect the payoffs of corporate securities and their equilibrium prices. For example, Kogan (2001) shows that irreversibility in firms' real investments can lead to time-varying stock risks, which can help to explain their returns. Zhang (2006) examines potential links between the time-varying risks from the real side and several empirical anomalies.

The empirical and theoretical work so far suggest that market frictions can be an important factor in determining asset prices. However, they are mainly indicative. The models and the phenomena they address tend to be quite specialized. A more general framework capable of providing both a qualitative characterization and a quantitative assessment of the importance of frictions on the market behaviour is still lacking. This in part reflects the complexity of the problem. In the presence of frictions, the behaviour of market participants and the interactions among them become much more involved, and the simple aggregation properties assumed in the neoclassical framework no longer hold. Whether a general theory will eventually emerge or we have to settle for a collection of specialized models to deal with each individual phenomenon remains unclear at this point.

Information Asymmetry

Information is a critical force driving financial markets. As is evident from (Eq. 4), it is the expectation of market participants of discounted future cash flows that determines asset prices. In general, information is asymmetric among different market participants. Under the extreme

situation when the market is competitive and sufficiently complete, the price system will be efficient in aggregating and revealing the information of all participants in the market, which gives a strong form of the EMH (see, for example, Grossman 1976; Milgrom and Stokey 1982). However, in the presence of frictions, in particular certain forms of market incompleteness, prices fail to be a sufficient static for the information in the market (see, for example, Grossman and Stiglitz 1980; Hellwig 1980; Diamond and Verrecchia 1981). While information asymmetry also contributes to the existence of frictions, most of the analysis takes certain form of frictions as given and examines the effect of information asymmetry.

Information asymmetry substantially enriches the possible behaviour of asset prices. In general, current prices do not reveal all the information in the market. This immediately implies that past prices or other public information can provide additional information over current prices (see, for example, Brown and Jennings 1989; Grundy and McNichols 1989). More importantly, under asymmetric information, the behaviour of market participants will depend not only on their own information but also on their perception of the information others may have. In an intertemporal equilibrium setting, Wang (1993) demonstrates that information asymmetry can have a broad impact on asset prices, ranging from increasing the risk premium and price volatility to generating rich patterns in return dynamics. Allen et al. (2006) further show that speculation on what others think may lead to price bubbles.

While information asymmetry increases the flexibility of the theory, its impact is harder to identify empirically as private information, by its nature, is mostly unobservable. By comparing price volatility on days when the stock market is open for trading with days when it is closed, French and Roll (1986) demonstrated convincingly the important role private information plays. Wang (1994) proposed using the joint behaviour of price and volume to examine the effect of information asymmetry (see also He and Wang 1995). Empirical work along this line, notably, Llorente et al. (2002), have found supporting evidence for this approach. Recently, more detailed data on individual investors' trading records has become available (for example, Odean 1998; Grinblatt and Keloharju 2000), which will allow more direct tests on the importance of information asymmetry. For example, following a segment of the market, Evans and Lyons (2002) find that order flow in the currency market contains significant amount of information.

Another challenge to the neoclassical theory is market crashes, that is, large price drops without significant macro news. If the prices before and after a crash both reflect the market's expectation of discounted future cash flows, either the discount rate or the expectation (or both) must have changed during crash. As discussed earlier, liquidity effect can cause the discount rate to vary abruptly, as shown by Huang and Wang (2006b). Alternatively, the market expectation can change, reflecting changes in the information it contains. In the absence of big exogenous news, this information must come from the private information investors already possess. Various models have been proposed to explain market crashes, including Grossman (1988), Gennotte and Leland (1990), Bikhchandani et al. (1992) and Romer (1993). These models typically allow both information asymmetry and market frictions, such as restrictions on what and how to trade, which prevent private information from being fully reflected in market prices. An unsettling issue for some of these models is the symmetry in large price movements they produce, that is, equal likelihood of market crashes and surges. Asymmetry in favour of crashes can arise when frictions of asymmetric nature are present, such as borrowing constraints (for example, Yuan 2005) and short-sale constraints (for example, Bai et al. 2006).

With regard to the impact of information asymmetry on return cross section, we face a similar situation as with frictions. The theory loses its tractability very quickly. Using a simple setting similar to that of the CAPM, Admati (1985) demonstrated that, under information asymmetry, the behaviour of equilibrium prices becomes very complex and sensitive to the information structure. We have not moved much beyond this point.

How information influences prices is a central issue in asset pricing. Existing work points to important channels for these influences, but the analysis is far from complete. On the one hand, the models so far are quite simplistic, especially in capturing the nature of information asymmetry in the market, and richer models are needed. On the other hand, even models with simple forms of information asymmetry are easily lost in their complexity. Both empirical and methodological breakthroughs are needed here.

Market Microstructure

Many frictions are endogenous. A lot of effort has been devoted to studying how certain frictions, in particular liquidity, are determined in the market through the actual trading process, which is also referred to as the market microstructure (Garman 1976). Despite its sophistication, the trading processes in the financial market are far more complex than what is assumed in most of the theoretical models, that is, through a Walrasian auction. The trading process also differs across different markets, ranging from over-the-counter markets and centralized exchanges with specialists to electronic limit order books, and constantly evolves over time. Several questions arise. How does a particular trading process influence the ease of trading or liquidity, investors' trading behaviour, and the properties of prices? How does it influence the efficiency of the market and overall asset valuation? What determines the form of the trading process in a given market and how it evolves?

A large body of work focuses on how market-makers, who provide liquidity by absorbing transitory order imbalances, influence effective trading costs and high-frequency price dynamics. Market-makers' behaviour depends on the costs they face, which have two components: the cost of holding an inventory and the cost of adverse selection when trading against better informed investors. Earlier analysis emphasized the former. Attributing the inventory cost to the risk in the value of inventory, Stoll (1979) showed that the effective trading cost, as measured by the bid-ask spread, increases with competitive market makers' risk aversion and the volatility of asset value (see also Amihud and Mendelson 1980). Roll (1984b) developed an empirical measure of the effective bid-ask spread and found it to be nontrivial for most individual stocks. Later attention has turned to the effect of adverse selection. Glosten and Milgrom (1985) showed how the existence of informed trades contributes to the bid-ask spread. Kyle (1985) demonstrated how an insider's strategic behaviour hinders the informational efficiency of the market and reduces its liquidity. Similar analysis have been carried out for markets organized as a limit order book, such as in Copeland and Galai (1983), Rock (1990) and Glosten (1994). High frequency data on quotes and trades made it possible for extensive studies of the behaviour of trading and prices in different markets, following the intuition developed in theory, including Glosten and Harris (1988), Hasbrouck (1991), Madhavan and Smidt (1993), Biais et al. (1995) and Lyons (1995). Imperfect competition among market makers also leads to additional complexity in the supply of liquidity (see, for example, Christie and Schultz 1994; Barclay et al. 1999; Wahal 1997; Ellis et al. 2002). Additional theoretical work has been directed at how market-makers behave strategically in their liquidity provision under different trading processes, notably Glosten (1989), Foucault (1999), Vayanos (1999), and Goettler et al. (2005).

Although most of the theoretical analysis on microstructure has focused on centralized markets, some considers the over-the-counter (OTC) markets, which by some measures are more common. Duffie et al. (2005) develop a search-based model for the OTC market. It was then applied to several markets such as securities borrowing (Duffie et al. 2005) and Treasury bonds (Vayanos and Weill 2006).

Market microstructure effects provide new insights on market behaviour at high frequency. For example, Admati and Pfleiderer (1988) and Foster and Viswanathan (1990) considered how traders' strategic behaviour in response to the liquidity in the market can help to explain intraday variations in trading volume and price volatility (see also Hong and Wang 2000, for alternative explanations). To the extent that market microstructure affects transactions costs in the market,

it also influences asset prices in general, as we discussed earlier (see also O'Hara 2003).

Many studies have also compared the different ways trading is organized, in particular how different market organizations may affect their liquidity provision and informational efficiency. For example, Copeland and Galai (1983) illustrated certain benefits of call auctions. Grossman (1992) examined the efficiency of upstairs market for block trades. Glosten (1994) discussed the advantage of an electronic limit order book. Seppi (1997) considered the impact of competition between a limit order book and a specialist when they coexist. Direct empirical comparisons of different trading mechanisms are difficult as they are usually adopted for different markets. But the general evidence is clear: market behaviour does vary with the actual trading process (for example, Amihud and Mendelson 1991; Ready 1999; Goldstein and Kavajecz 2000; Bessembinder 2003; Boehmer et al. 2005).

Although a lot has been learned about market microstructure, more remains to be learned. Many factors are at play and only a few are considered at a time, both theoretically and empirically. Their relative importance is hard to gauge empirically to allow for possible simplification. It remains a question why a given market is organized in a certain fashion. A better understanding of the precise nature and the magnitude of its impact on asset valuation and market efficiency is also needed.

From a broader perspective, there is also the question on the overall market structure (such as what securities are traded and why), which we may refer as market macrostructure. Most of the neoclassical theory takes it as given. But the dramatic evolution of the market, driven by a flood of innovations in finance and advances in technology and changes in the global economy, has forced researchers to think hard about this question. Some preliminary work has emerged in addressing this question. Allen and Gale (1988) consider the choice of firms in issuing securities when taking into account its impact on market structure and the resulting prices. Duffie and Rahi (1995) examine how exchanges decide on the derivative contracts to offer. Huang and Wang (1997) analyse how the introduction of new securities may influence the overall informational efficiency of the market. Of course, a significant number of financial transactions are carried out through financial intermediaries rather than in the form of financial securities. Allen and Gale (2004) further explore the interplay between the two. Despite the importance of this question, the work so far is extremely primitive and in many ways merely serves to keep the question in play.

Behavioural Finance

In the search for alternative explanations of asset pricing anomalies, attention has also turned to some of the basic assumptions of the neoclassical theory. The absence of arbitrage and the notion of efficient markets rely on the assumption that marginal investors in the market are not hindered by market frictions. The work on frictions has attempted to relax this assumption. Equilibrium models of asset pricing further adopt the assumption that average investors behave 'rationally'. However, the notion of rationality is an ambiguous one. Earlier models describe rationality in the form of expected utility, where the expectation is taken under the actual probability measure, for example, in the form of von Neumann and Morgenstern (1944). This implies that an investor's belief about market behaviour is consistent with its true behaviour. In addition, the utility function is assumed to depend only on the level of consumption. In a simple form, an investor's behaviour is described by the following expected utility

$$u_t(c_t) + \mathrm{E}_t[u_{t+1}(c_{t+1})] = u_t(c_t) + \sum_{\omega} p(\omega) u_{t+1}(c_{t+1}), \qquad (16)$$

where $p(\omega)$ is the actual probability for a future state ω and c_{t+1} is the level of future consumption. For simplicity, here we assume time-separable utility function and symmetric information. (In the case of asymmetric information, $p(\omega)$ becomes the probability conditional on the investor's information.) Since its justification is more normative than positive, this form of rationality has attracted many criticisms from very earlier on, notably, Allais (1953), Ellsberg (1961) and Kahneman and Tversky (1974). Deviations from

this simple form of rationality have gained prominence in various attempts to explain market anomalies. Since these explanations are mostly based on various assumptions on investor behaviour, this area of research has gained the name of behaviour finance.

Most of the evidence against the simple form of rationality is from laboratory experiments on human subjects with hypothetical prospects or small-stake choices. It was documented that subjects often fail to form objective and consistent probabilistic assessments, exhibiting patterns like overconfidence (for example, Fischhoff et. al. 1977; Weinstein 1980), belief perseverance, and anchoring (Kahneman and Tversky 1974). When facing gambles with stated probabilities, subjects' choices are incompatible with the expected utility, as documented by Kahneman and Tversky (1974). In addition, when confronted with outcomes with unknown probabilities, subjects' choices cannot be reconciled with a consistent probabilistic assessment of the possible outcomes (for example, Ellsberg 1961). Knight (1936) referred to this situation as uncertainty as opposed to risk, for which probabilities are known.

The richness in these behavioural variations, when used to describe investor behaviour, gives tremendous flexibility in providing possible explanations of asset price behaviour. For example, DeBondt and Thaler (1985) attribute the reversals in long-horizon market returns to investor overreaction. Using a version of prospect theory, in which investors exhibit loss aversion (that is, over weighting potential losses from a benchmark point over gains), Barberis et al. (2001) demonstrate that it can help to reconcile the high equity premium and price volatility with smooth consumption. Daniel et al. (2001) interpret the cross-sectional deviations in average equity returns from the CAPM, in particular the value and size premia, as a result of overconfidence in investors' interpretation of their private information. Models based on belief perseverance and representativeness (for example, Barberis et al. 1998), overconfidence (Daniel et al. 1998), and under-reaction to information (Hong and Stein 1999) have been used to explain anomalies like short-horizon return momentum, long-horizon return reversal, and post-earnings announcement drift. Liu et al. (2005) assume uncertainty aversion to reconcile the high premium of options paying off in rare events with their low probabilities seen in the data.

The experimental basis of behavioural assumptions raises the question of their relevance for actual individual behaviour in real economic decisions. As data on individual investments becomes available, more direct examination of their behaviour is possible. Investors are found to invest more in stocks they are familiar with (for example, French and Poterba 1991; Grinblatt and Keloharju 2001), to diversify naively (for example, Benartzi and Thaler 2001), to trade excessively (for example, Barber and Odean 2000), and to sell winners quickly while holding on to losers (for example, Odean 1998). These investment patterns are interpreted as being consistent with some of the behavioural assumptions. However, the presence of many other factors, ranging from taxes, information to portfolio considerations, leaves plenty of space for alternative interpretations.

Although the behavioural patterns explored in the literature are not fully described by the expected utility theory and are thus referred to as irrational, most of them can be captured by a more general form of rationality formulated by Savage (1954), which allows for subjective beliefs and state-dependent utility functions:

$$u_t(c_t) + \sum_{\omega} p^i(\omega) u_{t+1}\left(c_{t+1}, \omega\right), \qquad (17)$$

where $p^i(\omega)$ denotes the subjective probability of an investor i. The subjective expected utility theory in (Eq. 17) still exhibits a general form of consistency on behaviour but can accommodate rich variations in individual beliefs and preferences. In addition, within this more general notion of rationality, the distinction between beliefs and preferences become more of a formality. For example, a subjective expected utility function after the following transformation

$$u_t(c_t) + \sum_{\omega} p^i(\omega)\, u_{t+1}\left(c_{t+1}, \omega\right) = u_t(c_t) + \sum_{\omega} p(\omega)\, \tilde{u}_{t+1}\left(c_{t+1}, \omega\right),$$

where $\tilde{u}_{t+1}(c_{t+1}, \omega) = [p^i(\omega)/p(\omega)]u_{t+1}(c_{t+1}, \omega)$, becomes an expected utility function (under the true probability measure p) describing the same behaviour. From this point of view, we have three observations. First, many behavioural patterns can be obtained from state-dependent expected utility, a form of rationality slightly more general than that defined by state-independent expected utility. Second, with state dependent utility, many behaviour models are formally indistinguishable from those considered within the neoclassic framework. Third, without additional restrictions, the distinction between belief and preference is largely arbitrary.

As for the consumption-based CAPM with habits, behavioural models, as they stand now, also face major limitations. First, without additional discipline, the theory is simply too flexible. As the distance between assumption and result decreases, the multiplicity of potential explanations actually increases. Second, even taking the behavioural patterns at the individual level as given, it is less clear how they aggregate. Idiosyncratic biases at the individual level may well average out at the market level.

Another critical and perhaps more important issue is to what extent deviations in individual behaviour from rationality, even if they persist at the market level, influence asset prices. Take momentum as an example. If the predictability arises from the under-reaction of some investors to new information, investors who have information and capital, also referred to as arbitrageurs, should jump in to take advantage of the predictability until it disappears. As discussed in the section on market efficiency, two factors can hold back this market force, namely, risk and frictions. De Long et al. (1990) argued that irrationality can generate sufficient risk in the market to deter the arbitrageurs. But Sandroni (2000) demonstrated that, in a perfect market, investors acting on irrational beliefs do not survive in the long run (although their price impact may persist longer, as shown in Kogan et al. 2006). For risk to matter and the impact of irrational behaviour to persist, frictions or 'limits of arbitrage' are essential, a point emphasized by Shleifer and Vishny (1997). However, as discussed earlier in this section, in the presence of frictions various so-called anomalies can be accounted for without relying on additional behavioural assumptions. Faced with many competing and piece-wise 'theories', the challenge we face is to further pin down the actual causes of observed pricing patterns within a unified, and hopefully simple theory.

Corporate Finance

Guided by prices, the actual allocation of capital is achieved by transactions among market participants, mainly firms and households. Firms' financial behaviour is of particular importance as their main function is to create value from the existing capital, while households are the ultimate owner and beneficiaries. The pricing principles from the neoclassical finance have lent powerful tools for corporate and individual financial decision making. A better understanding of the financial behaviour of firms and individuals, which drives the demand and supply of assets, is also essential to our understanding of asset prices.

Corporate finance in the neoclassical theory began with the seminal work of Modigliani and Miller (1958, 1963). Using the principle of no arbitrage, they showed that in the absence of imperfections a firm's value depends only on its investment decisions. Financing and payout decisions merely determine how payoffs from investments are split between different claims associated with each financing vehicle – for example, equity and debt. The irrelevancy results of Modigliani and Miller (MM) clearly points to the areas where corporate finance matters. Much of the work since has focused on these areas where assumptions of MM are relaxed, in particular when frictions and information problems are important.

The information problems have been mostly framed in the interaction between a firm's insider who manages it and its outside investor who finances it. The insider can be an entrepreneur seeking outside capital or a manager running a mature public company. Two types of information problems were identified early on, namely, adverse selection and moral hazard. Leland and Pyle (1977)

and Ross (1977) examined how the adverse-selection problem influences the firm's financial decisions when firm insiders/managers know more about firm assets. In this case, firms' actions also serve as signals to outside investors, which will influence their perception of firm value. Viewing outside investors (for example, equity and debt holders) as the principal and managers as agents, Wilson (1968) and Ross (1973) considered the moral hazard problem when managers have more information on firm assets and their own actions. Based on this type of agency theory, Jensen and Meckling (1976) and Myers (1977) examined how firm value can be influenced by the conflicts between different stakeholders, that is, investors versus managers and shareholders versus bondholders.

Recent developments in corporate finance have followed this theme. Corporate behaviour was often viewed as a manifestation of these conflicts. In order to turn this general perspective into testable theories, more structure is needed with regard to the nature and the magnitude of these imperfections. In this regard, more guidance from the data becomes critical. Our discussion starts with how a firm chooses its financing or capital structure, the focus of neoclassical theory. We then turn to how inefficiencies in financing caused by frictions and information problems influence a firm's investments. Finally, we consider the issues concerning corporate control, which looks at the problems in corporate finance from a more fundamental perspective.

Financing

Financially, a firm is about how to raise capital and how to use it. The two questions are obviously intertwined. Under MM, the two become independent. A firm's overall cost of capital, that is, the valuation of its assets, is not affected by how it is financed. An important friction omitted in this irrelevancy result is taxes. With different tax treatments on debt and equity financing, different choices of capital structure will affect the firm's tax liability and naturally its value. For example, when interest payments on corporate debt are excluded from corporate taxes, the firm can pass on higher returns to its investors by substituting debt for equity. This tax arbitrage, however, has its barbs. First, it does not account for investors' personal taxes. Miller (1977a, b) showed that, as investors of different tax clienteles settle for different mixes of debt and equity, an equilibrium is reached when marginal investors are indifferent between the two, which determines the total amount of debt and equity but not for individual firms as their securities are substitutes. Second, it does not take into account the potential cost of using debt, which can lead to bankruptcy. When the effects of personal and corporate taxes are different (for example, when securities of different firms are not perfect substitutes) and bankruptcy is costly, we have the so-called 'trade-off' theory of capital structure. Each firm is trading off the tax benefits of debt and the bankruptcy costs.

How significant these benefits and costs are remains an empirical question. Data seems to suggest that they are important. For example, Graham (2000) estimates that the effective tax rate paid by marginal investors on debt is significantly higher than that on equity, suggesting a large benefit of debt finance. The cost of bankruptcy has several sources, the direct cost of bankruptcy process (for example, Weiss and Wruck 1998) and the indirect cost of financial distress such as conflicts between different stakeholders (for example, Asquith and Wizman 1990), loss of business and financial counter parties (for example, Maksimovic and Titman 1991), pressure from competitors (for example, Chevalier 1995). A more recent study by Andrade and Kaplan (1998) estimates costs of financial distress to be in the range of 10–20 per cent of the firm value prior to distress. *Ex post* the cost of this size may seem modest since on an *ex ante* basis one has to factor in its probability. However, the fact that these costs tend to have large negative beta (negatively correlated with the market) may imply that their present value is non-trivial (for example, Almeida and Philippon 2006).

The trade-off theory has several implications. First, each firm should have an optimal capital structure, which depends on its tax status and cost of financial distress. Although the theory does not fully specify what determines these two factors, different proxies were used empirically.

For example, firms with higher business risk and more intangible assets were associated with higher distress costs and thus a low debt–asset ratio. Many empirical studies have found positive evidence on the link between these proxies and the capital structure, such as Auerbach (1985), Titman and Wessels (1988) and Rajan and Zingales (1995). But the evidence is not uniformly supportive. Wald (1999) finds that profitability has a strong negative relation with debt–asset ratios, while the trade-off theory would imply that more profitable firms should use more debt to shield their income. Second, if adjustment is costly, a firm's capital structure will be away from its optimum most of the time but always evolves towards it. As a result, the firm is more likely to issue debt when below the target and equity when above. Earlier tests found this prediction to be consistent with the data (for example, Taggart 1977; Auerbach 1985), but more recent tests have found mixed evidence (for example, Hovakimian et al. 2001). The trade-off theory is intuitive and enjoys partial empirical success, but still leaves some gaps. In particular, the significant costs of financial distress need both theoretical and empirical justification.

Based on patterns in firms' financing choices, Myers and Majluf (1984) propose the pecking-order theory of capital structure. It starts with the premise that outsider investors have less information about a firm's use of capital. Thus, they face an adverse selection problem and will on average undervalue new shares. Equity becomes more costly as a financing vehicle than debt. This simple theory yields several predictions: (*a*) firms prefer internal to external finance and debt to equity finance; (*b*) the market reacts negatively to new share issues; (*c*) dividends are persistent; (*d*) a firm's debt–asset ratio changes with its cumulative needs for external financing. Heuristic empirical observations are surprisingly compatible with the pecking-order theory. But more extensive tests reveal some inconsistencies. For example, Jung et al. (1996) and Fama and French (2002) have found that small-growth firms rely heavily on equity financing. Although the pecking-order theory is very much in the spirit of Ross (1977), it relies on simplifying assumptions. In particular, no optimal contracting is considered by allowing for more complex forms of financing and incentives to resolve the information problem.

The agency theory of capital structure focuses on the conflict of interest between managers and shareholders. This is different from the trade-off and pecking-order theories, in which managers act on behalf of current shareholders. When their information and actions are not fully observable to outsiders, which may include current shareholders, managers can benefit themselves at the cost of shareholders. When incentives through contracting fail to fully mitigate this conflict, capital structure will be influenced by investors' efforts to contain managers. Following Jensen (1986), different variations of the agency theory have been proposed, notably Harris and Raviv (1993), Stulz (1990), and Zwiebel (1996). For example, Jensen (1986) argued that debt helps to get cash out of managers' hands and is thus preferred to outside equity before bankruptcy becomes important. Although empirically leverage does curb investments (for example, Lang et al. 1996), new debt issues do not seem to increase firm value (Eckbo 1986).

Factors like taxes, cost of financial distress, information and agency problems do matter for firms' financing decisions, as the empirical evidence suggests. But each of the theories captures only part of the picture. They are also mostly partial equilibrium by nature, taking certain aspects of the problem as given such as the contracting environment and firms' investment opportunities. A more integrated and empirically refutable theory would be desirable. On the empirical side, it remains a challenge to reconcile the financing patterns found under certain circumstances with the lack of a link between taxes, financing and market value documented in Fama and French (1998) over a large sample of firms.

Investments

Clearly, the forces driving a firm's financing choices also influence its investments, that is, the use of capital. We have identified at least two channels. The first channel is simply through the cost of capital. In the case of trade-off theory, for

example, a firm's cost of capital varies with its capital structure and so will its investments. The second channel is through the behaviour of managers, who make investment decision in response to the incentives they face, which are also related to the firm's financing choices.

The direct effect of cost of capital has found supportive evidence. For example, using a structural model to calibrate firms' investment opportunities, Hennessy (2004) finds that a high debt level curbs investments. In the state of distress, firms also cut down their investments, as documented in many studies, including Chevalier (1995b), Phillips (1995) and Zingales (1998).

The agency effect has attracted more interest. A variety of private benefits of managers were suggested, ranging from empire building (Williamson, 1964; Jensen, 1986) and career considerations (Narayanan, 1985; Holmstrom, 1999) to inertia (for example, Bertrand and Mullainathan 2003). Misalignment between managers' and shareholders' interests will lead to suboptimal investment decisions. A simple prediction of this argument is that firms with more free cash in hand will make more and less desirable investments. Broad empirical evidence was found to be consistent with this prediction, such as Fazzari et al. (1988), Hoshi et al. (1991), and Gilchrist and Himmelberg (1995). One challenge in establishing the empirical link is the problem of endogeneity. For example, a firm's free cash is endogenous and may vary with its investments opportunities. Several studies have used 'natural experiments' to avoid the endogeneity issue. For example, Blanchard et al. (1994) find that firms' acquisition activities increase after receiving cash windfalls from legal settlements unrelated to their business.

A positive correlation between cash and investments does not prove the agency theory. It can also be consistent with the pecking-order theory. More cash relaxes the capital constraint imposed by high cost of external financing. The question is whether free cash flow leads to negative net present value (NPV) investments. Evidence such as the negative price reaction to new equity issues (for example, Asquith and Mullins 1986) may suggest so. However, more direct evidence indicates otherwise. For example, McConnell and Muscarella (1985) documented positive market reactions to firms' capital expenditure announcements.

Firms' investment decisions are of central importance to finance. Frictions and information problems imply inefficient use of capital. Extensive evidence is suggestive of such inefficiencies, but it is far from definitive. A comprehensive empirical evaluation of the extent and the magnitude of these efficiencies and the potential forces driving them is not available yet.

F

Corporate Control and Governance

Most of the new theories in corporate finance take as given the means different parties use to resolve conflicts. For example, in the pecking-order theory or the agency theory of financing, the mixture of debt and equity is the tool available to balance the interests of managers and outside investors. But a whole set of devices can be utilized to resolve their conflicts, including incentive contracts for managers and a rich set of corporate securities beyond equity and bonds. It makes sense to think at a deeper level about the economic structure of a firm and its resulting behaviour.

Built on the ideas of Coase (1937), Grossman and Hart (1986) proposed the idea that a firm is defined by the allocation of control rights over its assets, the rights to utilize these assets. In such a setting, conflicts among different stakeholders are resolved by optimal allocation of control rights rather than extensive contracting, which is assumed to be infeasible with hard-to-specify future contingencies. Such an allocation will then determine how the firm behaves, including its investment decisions and financing arrangements. It will also determine how it is governed – for example, who takes control and when. A collection of theories on corporate behaviour was developed under this setting.

Aghion and Bolton (1992) considered the financing problem of an entrepreneur who also enjoys private benefits from running his firm (see also Hart and Moore 1998). The optimal structure of the firm would be for him to retain the control rights of the firm (so he can enjoy the private benefits) while selling cash flow claims to

outside investors. This looks very much like a mixture of equity and debt financing, except that now it is the outcome of optimal corporate control. If embedded in an intertemporal environment with uncertainty, the model also lead to implications on the dynamics of the firm's financing and investments. By looking at venture capital investments in start-up companies, Gompers (1995) and Kaplan and Stromberg (2001) have found patterns compatible with the model's predictions. This model, however, captures mostly inside equity and is less descriptive of large public firms, which involves mostly equity held by outsiders. Fluck (1998) and Myers (2000) have considered models for outside equity financing. Within a similar framework, Grossman and Hart (1980, 1988) analyse the market for corporate control in the form of takeovers (see also Harris and Raviv 1988) when shareholdings are diverse. Aghion and Tirole (1997) examined issues concerning corporate governance, such as the role of corporate boards, which act as shareholders' representative in exercising their control rights.

Models of incomplete contracting capture some salient features of firms and attempt to examine corporate finance issues from a more basic and integrated perspective. But they are highly simplified. Their predications are mostly qualitative and dependent on deeper parameters, such as what can or cannot be contracted. The fact that these parameters are hard to observe make it hard to empirically test the models.

Another approach is to consider the firm as a full contract among its stakeholders, including managers and outside investors, very much in the spirit of Leland and Pyle (1977) and Ross (1977) (see also Townsend 1978; Gale and Hellwig 1985). For example, Gertler (1992), Clementi and Hopenhayn (2006), and DeMarzo and Fishman (2007) examine optimal contracts between investors and a manager to induce optimal investments. Atkeson and Cole (2005) consider the optimal financing contract in the presence of agency problems and manager risk aversion. In contrast to the assumptions in the models based on incomplete contracts, this approach explores what optimal contracting can achieve. As shown in Dybvig and Zender (1991), under certain circumstances optimal contracting can largely resolve the information problems between managers and shareholders.

The full contracting approach avoids some of the arbitrariness in the theory of incomplete contracts. But it has its own challenges. It is quite limited in describing large public companies, which involves a large number of stakeholders, including a hierarchy of managers and a diverse set of investors. Its predictions depend on the assumptions about other frictions such as verification and enforcement costs. Realistic assumptions about these frictions are also hard to pin down. This also leads to the question of the robustness of contractual arrangements from the models.

Conclusion

Developments in finance since the mid-1980s have expanded the success of neoclassical theory, especially in the area of arbitrage pricing, as well as its boundaries. Extensive and more rigorous empirical analysis has exposed the limitations of the simple asset pricing models and the simplistic notion of market efficiency. The fact that we still don't have a satisfactory notion of risk and can't explain movements in asset prices after the fact clearly suggests the need to enrich our theory. Imperfections such as frictions and information asymmetry are part of the market reality and should be incorporated. They can very much enhance our understanding of market participants' behaviour and its impact on the market itself. A rich set of models, accompanied by empirical work, has been explored to explain the observed patterns in the financial market and in corporate finance. Liquidity and agency problems have been identified as manifestations of imperfections in the market and corporate contexts and useful lenses through which to examine their behaviour.

An unavoidable challenge in modelling imperfections is that they come in all shapes and sizes, and their impact is in general complex. Empirical evaluation of the significance of various imperfections is very much needed to arrive at a unified framework, synthesizing the important intuition from the collection of specialized models we

have. After carefully collecting and studying the pieces and parts, we may be able to hope for a more general theory of finance.

See Also

- ▶ Arbitrage Pricing Theory
- ▶ Capital Asset Pricing Model
- ▶ Corporate Governance
- ▶ Dividend Policy
- ▶ Finance
- ▶ Risk

Bibliography

Abel, A.B.. 1990. Asset prices under habit formation and catching up with the Joneses. *American Economic Review* 80: 38–42.

Acharya, V., and L. Pedersen. 2005. Asset pricing with liquidity risk. *Journal of Financial Economics* 77: 385–410.

Admati, A.R. 1985. A noisy rational expectations equilibrium for multiple asset securities markets. *Econometrica* 53: 629–657.

Admati, A.R., and P. Pfleiderer. 1988. A theory of intraday patterns: Volume and price variability. *Review of Financial Studies* 1: 3–40.

Adrian, T., and J. V. Rosenberg. 2006. Stock returns and volatility: Pricing the short-run and long-run components of market risk. Staff Report No. 254. Federal Reserve Bank of New York.

Aghion, P., and P. Bolton. 1992. An incomplete contracts approach to financial contracting. *Review of Economic Studies* 59: 473–494.

Aghion, P., and J. Tirole. 1997. Formal and real authority in organizations. *Journal of Political Economy* 105: 1–29.

Ahn, D.H., R.F. Dittmar, and A.R. Gallant. 2002. Quadratic term structure models: Theory and evidence. *Review of Financial Studies* 16: 243–288.

Ait-Sahalia, Y. 1996. Nonparametric pricing of interest rate derivative securities. *Econometrica* 64: 527–560.

Allais, M. 1953. Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'École américaine. *Econometrica* 21: 503–546.

Allen, F., and D. Gale. 1988. Optimal security design. *Review of Financial Studies* 1: 229–263.

Allen, F., and D. Gale. 1994. Limited market participation and volatility of asset prices. *American Economic Review* 84: 933–955.

Allen, F., and D. Gale. 2004. Financial intermediaries and markets. *Econometrica* 72: 1023–1061.

Allen, F., S. Morris, and H.S. Shin. 2006. Beauty contests and iterated expectations in asset markets. *Review of Financial Studies* 19: 719–752.

Almeida, H., and T. Philippon. 2006. The risk-adjusted cost of financial distress. Working paper. New York University.

Amihud, Y., and H. Mendelson. 1980. Dealership market: Market making with inventory. *Journal of Financial Economics* 8: 31–53.

Amihud, Y., and H. Mendelson. 1986. Asset pricing and the bid–ask spread. *Journal of Financial Economics* 17: 223–249.

Amihud, Y., and H. Mendelson. 1991. Volatility, efficiency, and trading: Evidence from the Japanese stock market. *Journal of Finance* 46: 1765–1789.

Amin, K.I. 1993. Jump diffusion option valuation in discrete time. *Journal of Finance* 48: 1833–1863.

Anderson, R.W., and S. Sundaresan. 1996. Design and valuation of debt contracts. *Review of Financial Studies* 9: 37–68.

Andrade, G., and S.N. Kaplan. 1998. How costly is financial (not economic) distress? Evidence from highly leveraged transactions that become distressed. *Journal of Finance* 53: 1443–1493.

Ang, A., R.J. Hodrick, Y. Xing, and X. Zhang. 2006. The cross-section of volatility and expected returns. *Journal of Finance* 51: 259–299.

Asquith, P., and D.W. Mullins. 1986. Equity issues and offering dilution. *Journal of Financial Economics* 15: 61–89.

Asquith, P., and T.A. Wizman. 1990. Event risk, covenants, and bondholder returns in leveraged buyouts. *Journal of Financial Economics* 27: 195–213.

Atkeson, A., and H. Harold. 2005. A dynamic theory of optimal capital structure and executive compensation. Working Paper No. 11083. Cambridge, MA: NBER.

Auerbach, A.S. 1985. Real determinants of corporate leverage. In *Corporate capital structure in the United States*, ed. B.M. Friedman. Chicago: University of Chicago Press.

Bai, Y., E. C. Chang, and J. Wang. 2006. Asset prices and short-sale constraints. Working Paper. Massachusetts Institute of Technology.

Bakshi, G., C. Cao, and Z. Chen. 1997. Empirical performance of alternative option pricing models. *Journal of Finance* 52: 2003–2049.

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

Bansal, R., R.F. Dittmar, and C.T. Lundblad. 2005. Consumption, dividend, and the cross section of equity returns. *Journal of Finance* 60: 1639–1672.

Bansal, R., and A. Yaron. 2004. Risks for the long run: A potential resolution of asset pricing puzzles. *Journal of Finance* 59: 1481–1510.

Banz, R.W. 1981. The relationship between return and market value of common stocks. *Journal of Financial Economics* 9: 3–18.

Barber, B.M., and T. Odean. 2000. Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance* 55: 773–806.

F

Barberis, N., M. Huang, and T. Santos. 2001. Prospect theory and asset prices. *Quarterly Journal of Economics* 116: 1–53.

Barberis, N., A. Shleifer, and R. Vishny. 1998. A model of investor sentiment. *Journal of Financial Economics* 49: 307–343.

Barclay, M.J., W.G. Christie, J.H. Harris, E. Kandel, et al. 1999. The effects of market reform on the trading costs and depths of Nasdaq stocks. *Journal of Finance* 54: 1–34.

Basak, S. 1995. A general equilibrium model of portfolio insurance. *Review of Financial Studies* 8: 1059–1090.

Basu, S. 1983. The relationship between earnings' yield, market value and return for NYSE common stocks: Further evidence. *Journal of Financial Economics* 12: 129–156.

Bates, D.S. 1996. Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche mark options. *Review of Financial Studies* 9: 69–107.

Bates, D.S. 2000. Post-'87 crash fears in the S&P 500 futures option market. *Journal of Econometrics* 94: 181–238.

Bessembinder, H. 2003. Trade execution costs and market quality after decimalization. *Journal of Financial and Quantitative Analysis* 38: 747–777.

Benartzi, S., and R.H. Thaler. 2001. Naive diversification strategies in defined contribution savings plans. *American Economic Review* 91: 79–98.

Berk, J.B., R.C. Green, and V. Naik. 1999. Optimal investment, growth options and security returns. *Journal of Finance* 54: 1553–1607.

Bernard, V.L., and J.K. Thomas. 1989. Post-earningsannouncement drift: Delayed price response or risk premium? *Journal of Accounting Research* 27: 1–36.

Bertrand, M., and S. Mullainathan. 2003. Enjoying the quiet life? Corporate governance and managerial preferences. *Journal of Political Economy* 111: 1043–1075.

Biais, B., P. Hillion, and C. Spatt. 1995. An empirical analysis of the limit order book and order flow in the Paris bourse. *Journal of Finance* 50: 1655–1689.

Bikhchandani, S., D. Hirshleifer, and I. Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy* 100: 992–1026.

Black, F., and J.C. Cox. 1976. Valuing corporate securities: Some effects of bond indenture provisions. *Journal of Finance* 31: 351–367.

Black, F., M.C. Jensen, and M. Scholes. 1972. The capital asset pricing model: Some empirical tests. In *Studies in the theory of capital markets*, ed. M.C. Jensen. New York: Praeger.

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.

Blanchard, O.J., F. Lopez-de-Silanes, and A. Shleifer. 1994. What do firms do with cash windfalls? *Journal of Financial Economics* 36: 337–360.

Boehmer, E., G. Saar, and L. Yu. 2005. Lifting the veil: An analysis of pre-trade transparency at the NYSE. *Journal of Finance* 60: 783–815.

Brav, A., G.M. Constantinides, and C.C. Geczy. 2002. Asset pricing with heterogeneous consumers and limited participation: Empirical evidence. *Journal of Political Economy* 110: 793–842.

Breeden, D.T. 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7: 265–296.

Breeden, D.T. 1980. Consumption risk in futures markets. *Journal of Finance* 35: 503–520.

Brennan, M.J., and E.S. Schwartz. 1979. A continuous time approach to the pricing of bonds. *Journal of Banking and Finance* 3: 133–155.

Brennan, M.J., and A. Subrahmanyam. 1996. Market microstructure and asset pricing: On the compensation for illiquidity in stock returns. *Journal of Financial Economics* 41: 341–364.

Brown, D.P., and R.H. Jennings. 1989. On technical analysis. *Review of Financial Studies* 2: 527–551.

Brown, R.H., and S. M. Schaefer. 1994. The term structure of real interest rates and the Cox Ingersoll and Ross model. *Journal of Financial Economics* 35: 3–42. Online available at http://econpapers.repec.org/article/eeejfinec/

Brown, S.J., and P.H. Dybvig. 1986. The empirical implications of the Cox, Ingersoll Ross theory of the term structure of interest rates. *Journal of Finance* 41: 617–630.

Campbell, J.Y., and J. Cochrane. 1999. By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy* 107: 205–251.

Campbell, J.Y., S.J. Grossman, and J. Wang. 1993. Trading volume and serial correlation in stock returns. *Quarterly Journal of Economics* 108: 905–939.

Campbell, J.Y., and R.J. Shiller. 1988. The dividend–price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1: 195–228.

Chan, K.C., G.A. Karolyi, F.A. Longstaff, and A.B. Sanders. 1992. An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance* 47: 1209–1227.

Chan, L.K., and J. Lakonishok. 1995. The behavior of stock prices around institutional trades. *Journal of Finance* 50: 1147–1174.

Chan, Y.L., and L. Kogan. 2002. Catching up with the Joneses: Heterogeneous preferences and the dynamics of asset prices. *Journal of Political Economy* 110: 1255–1285.

Chen, J., H. Hong, and J. Stein. 2002. Breadth of ownership and stock returns. *Journal of Financial Economics* 66: 171–205.

Chen, N., R. Roll, and S.A. Ross. 1986. Economic forces and the stock market. *Journal of Business* 59: 383–404.

Chen, R., and L. Scott. 1992. Pricing interest rate options in a two-factor Cox–Ingersoll–Ross model of the term structure. *Review of Financial Studies* 5: 613–636.

Chevalier, J.A. 1995. Capital structure and product market competition: Empirical evidence from the supermarket industry. *American Economic Review* 85: 415–435.

Chordia, T., R. Roll, and A. Subrahmanyam. 2000. Commonality in liquidity. *Journal of Financial Economics* 56: 3–28.

Christie, W.G., and P.H. Schultz. 1994. Why do Nasdaq market makers avoid odd-eighth quotes? *Journal of Finance* 49: 1813–1840.

Clementi, G.L., and H.A. Hopenhayn. 2006. A theory of financing constraints and firm dynamics. *Quarterly Journal of Economics* 121: 229–265.

Coase, R.H. 1937. The nature of the firm. *Economica* 4: 386–405.

Connor, G., and R.A. Korajczyk. 1988. Risk and return in an equilibrium APT: Application of a new test methodology. *Journal of Financial Economics* 21: 255–290.

Constantinides, G.M. 1986. Capital market equilibrium with transaction costs. *Journal of Political Economy* 94: 842–862.

Constantinides, G.M. 1990. Habit formation: A resolution of the equity premium puzzle. *Journal of Political Economy* 98: 519–543.

Constantinides, G.M. 1992. A theory of the nominal term structure of interest rates. *Review of Financial Studies* 5: 531–552.

Constantinides, G.M., J.B. Donaldson, and R. Mehra. 2002. Junior can't borrow: A new perspective on the equity premium puzzle. *Quarterly Journal of Economics* 117: 269–296.

Constantinides, G.M., and D. Duffie. 1996. Asset pricing with heterogeneous consumers. *Journal of Political Economy* 104: 219–240.

Copeland, T.E., and D. Galai. 1983. Information effects and the bid–ask spread. *Journal of Finance* 38: 1457–1469.

Courtadon, G. 1982. The pricing of options on default-free bonds. *Journal of Financial and Quantitative Analysis* 17: 75–100.

Coval, J., and E. Stafford. 2007. Asset fire sales (and purchases) in equity markets. *Journal of Financial Economics* 86: 479–512.

Cox, J.C., J.E. Ingersoll, and S.A. Ross. 1980. An analysis of variable rate loan contracts. *Journal of Finance* 35: 389–403.

Cox, J.C., J.E. Ingersoll, and S.A. Ross. 1985a. An intertemporal general equilibrium model of asset prices. *Econometrica* 53: 363–384.

Cox, J.C., J.E. Ingersoll, and S.A. Ross. 1985b. A theory of the term structure of interest rates. *Econometrica* 53: 385–407.

Cox, J.C., and S.A. Ross. 1976a. A survey of some new results in financial option pricing theory. *Journal of Finance* 31: 383–402.

Cox, J.C., and S.A. Ross. 1976b. The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3: 145–166.

Cox, J.C., S.A. Ross, and M. Rubinstein. 1979. Option pricing: A simplified approach. *Journal of Financial Economics* 7: 229–263.

Dai, Q., and K.J. Singleton. 2000. Specification analysis of affine term structure models. *Journal of Finance* 55: 1943–1978.

Daniel, K.D., D. Hirshleifer, and A. Subrahmanyam. 1998. Investor psychology and security market under- and over-reactions. *Journal of Finance* 53: 1839–1886.

Daniel, K.D., D. Hirshleifer, and A. Subrahmanyam. 2001. Overconfidence, arbitrage, and equilibrium asset pricing. *Journal of Finance* 56: 921–965.

De Long, J.B., A. Shleifer, L.H. Summers, and R.J. Waldman. 1990. Noise trader risk in financial markets. *Journal of Political Economy* 98: 703–738.

DeBondt, W.F.M., and R. Thaler. 1985. Does the stock market overreact? *Journal of Finance* 40: 793–805.

DeMarzo, P.M., and M.J. Fishman. 2007. Agency and optimal investment dynamics. *Review of Financial Studies* 20: 151–188.

Diamond, D.W., and R.E. Verrecchia. 1981. Information aggregation in a noisy expectations economy. *Journal of Financial Economics* 9: 221–235.

Diether, K.B., C.J. Malloy, and A. Scherbina. 2002. Differences in opinion and the cross section of stock returns. *Journal of Finance* 57: 2113–2141.

Duffie, D. 1996. Special repo rates. *Journal of Finance* 51: 493–526.

Duffie, D., N. Garleanu, and L.H. Pedersen. 2002. Securities lending, shorting, and pricing. *Journal of Financial Economics* 66: 307–339.

Duffie, D., N. Garleanu, and L.H. Pedersen. 2005. Over-the-counter markets. *Econometrica* 73: 1815–1847.

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

Duffie, D., and R. Rahi. 1995. Financial market innovation and security design: An introduction. *Journal of Economic Theory* 65: 1–42.

Duffie, D., and K.J. Singleton. 1999. Modeling term structures of defaultable bonds. *Review of Financial Studies* 12: 687–720.

Dumas, B. 1989. Two-person dynamic equilibrium in the capital market. *Review of Financial Studies* 2: 157–188.

Dybvig, P.H., and S.A. Ross. 1985. Yes, the APT is testable. *Journal of Finance* 40: 1173–1188.

Dybvig, P.H., and J.F. Zender. 1991. Capital structure and dividend irrelevance with asymmetric information. *Review of Financial Studies* 4: 201–219.

Eckbo, E.B. 1986. Mergers and the market for corporate control: The Canadian evidence. *Canadian Journal of Economics* 19: 236–260.

Ellis, K., R. Michaely, and M. O'Hara. 2002. The making of a dealer market: From entry to equilibrium in the trading of Nasdaq stocks. *Journal of Finance* 57: 2289–2316.

Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* 75: 643–669.

Epstein, L.G., and S.E. Zin. 1991. Substitution, risk aversion, and temporal behavior of consumption and asset returns: An empirical analysis. *Journal of Political Economy* 99: 263–286.

Evans, M.D.D., and R.K. Lyons. 2002. Order flow and exchange rate dynamics. *Journal of Political Economy* 110: 170–180.

Fama, E.F. 1970. Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25: 383–417.

Fama, E.F., and K.R. French. 1988. Permanent and temporary components of stock prices. *Journal of Political Economy* 96: 246–273.

Fama, E.F., and K.R. French. 1992. The cross-section of expected stock returns. *Journal of Finance* 47: 427–465.

Fama, E.F., and K.R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33: 5–56.

Fama, E.F., and K.R. French. 1998. Value versus growth: The international evidence. *Journal of Finance* 53: 1975–1979.

Fama, E.F., and K.R. French. 2002. Testing trade-off and pecking order predictions about dividends and debt. *Review of Financial Studies* 15: 1–33.

Fama, E.F., and J.D. MacBeth. 1973. Risk, return and equilibrium: Empirical tests. *Journal of Political Economy* 81: 607–636.

Fazzari, S.M., R.G. Hubbard, and B.C. Petersen. 1988. Financing constraints and corporate investment. *Brookings Papers on Economic Activity* 1: 141–195.

Fischhoff, B., P. Slovic, and S. Lichtenstein. 1977. Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology* 3: 552–564.

Fluck, Z. 1998. Optimal financial contracting: Debt versus outside equity. *Review of Financial Studies* 11: 383–418.

Foster, F.D., and S. Viswanathan. 1990. A theory of the intraday variations in volume, variance, and trading costs in securities markets. *Review of Financial Studies* 3: 593–624.

Foucault, T. 1999. Order flow composition and trading costs in a dynamic limit order market. *Journal of Financial Markets* 2: 99–134.

French, K.R., and J.M. Poterba. 1991. Investor diversification and international equity markets. *American Economic Review* 81: 222–226.

French, K.R., and R. Roll. 1986. Stock return variances: The arrival of information and reaction of traders. *Journal of Financial Economics* 17: 5–26.

Gale, D., and M. Hellwig. 1985. Incentive-compatible debt contracts: The one-period problem. *Review of Economic Studies* 52: 647–663.

Garman, M.B. 1976. Market microstructure. *Journal of Financial Economics* 3: 257–275.

Gennotte, G., and H. Leland. 1990. Market liquidity, hedging and crashes. *American Economic Review* 80: 999–1021.

Gertler, M. 1992. Financial capacity and output fluctuations in an economy with multi-period financial relationships. *Review of Economic Studies* 59: 455–472.

Gervais, S., R. Kaniel, and D.H. Mingelgrin. 2001. The high-volume return premium. *Journal of Finance* 56: 877–919.

Gibbons, M., and K. Ramaswamy. 1993. A test of the Cox, Ingersoll and Ross model of the term structure. *Review of Financial Studies* 6: 619–658.

Gilchrist, S., and C.P. Himmelberg. 1995. Evidence on the role of cash flow for investment. *Journal of Monetary Economics* 36: 531–572.

Glosten, L.R. 1989. Insider trading, liquidity, and the role of the monopolist specialist. *Journal of Business* 62: 211–235.

Glosten, L.R. 1994. Is the electronic open limit order book inevitable? *Journal of Finance* 49: 1127–1161.

Glosten, L.R., and L.E. Harris. 1988. Estimating the components of the bid–ask spread. *Journal of Financial Economics* 21: 123–142.

Glosten, L.R., and P.R. Milgrom. 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics* 14: 71–100.

Goettler, R.L., C.A. Parlour, and U. Rajan. 2005. Equilibrium in a dynamic limit order market. *Journal of Finance* 60: 2149–2192.

Goldstein, M.A., and K.A. Kavajecz. 2000. Eighths, sixteenths and market depth: Changes in tick size and liquidity provision on the NYSE. *Journal of Financial Economics* 56: 125–149.

Goldstein, R.S. 2000. The term structure of interest rates as a random field. *Review of Financial Studies* 13: 365–384.

Gomes, J.F., L. Kogan, and L. Zhang. 2003. Equilibrium cross section of returns. *Journal of Political Economy* 111: 693–732.

Gompers, P.A. 1995. Optimal investment, monitoring, and the staging of venture capital. *Journal of Finance* 50: 1461–1489.

Graham, J.R. 2000. How big are the tax benefits of debt? *Journal of Finance* 55: 1901–1941.

Gray, S.F. 1996. Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics* 42: 27–62.

Grinblatt, M., and M. Keloharju. 2000. The investment behavior and performance of various investor types: A study of Finland's unique data set. *Journal of Financial Economics* 55: 43–67.

Grinblatt, M., and M. Keloharju. 2001. What makes investors trade? *Journal of Finance* 56: 589–616.

Grossman, S.J. 1976. On the efficiency of competitive stock markets where traders have diverse information. *Journal of Finance* 31: 573–585.

Grossman, S.J. 1988. An analysis of the implications for stock and futures price volatility of program trading and dynamic hedging strategies. *Journal of Business* 61: 275–298.

Grossman, S.J. 1992. The informational role of upstairs and downstairs trading. *Journal of Business* 65: 509–528.

Grossman, S.J., and O.D. Hart. 1980. Takeover bids, the free-rider problem, and the theory of the corporation. *Bell Journal of Economics* 11: 42–64.

Grossman, S.J., and O.D. Hart. 1986. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy* 94: 691–719.

Grossman, S.J., and O.D. Hart. 1988. One share-one vote and the market for corporate control. *Journal of Financial Economics* 20: 175–202.

Grossman, S.J., and M.H. Miller. 1988. Liquidity and market structure. *Journal of Finance* 43: 617–637.

Grossman, S.J., and R.J. Shiller. 1981. The determinants of the variability of stock market prices. *American Economic Review* 71: 222–227.

Grossman, S.J., and J.E. Stiglitz. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70: 393–408.

Grossman, S.J., and J.L. Vila. 1992. Optimal dynamic trading with leverage constraints. *Journal of Financial and Quantitative Analysis* 27: 151–168.

Grossman, S.J., and Z. Zhou. 1996. Equilibrium analysis of portfolio insurance. *Journal of Finance* 51: 1379–1403.

Grundy, B., and M. McNichols. 1989. Trade and revelation of information through prices and direct disclosure. *Review of Financial Studies* 2: 495–526.

Hamilton, J.D. 1988. Rational-expectations econometric analysis of changes in regime: An investigation of the term structure of interest rates. *Journal of Economic Dynamics and Control* 12: 385–423.

Hansen, L.P., and R. Jagannathan. 1991. Implications of security market data for models of dynamic economies. *Journal of Political Economy* 99: 225–262.

Hansen, L.P., and K.J. Singleton. 1983. Stochastic consumption, risk aversion and the temporal behavior of asset returns. *Journal of Political Economy* 91: 249–268.

Harris, M., and A. Raviv. 1988. Corporate control contests and capital structure. *Journal of Financial Economics* 20: 55–86.

Harris, M., and A. Raviv. 1993. Differences of opinion make a horse race. *Review of Financial Studies* 6: 473–506.

Harrison, J.M., and D. Kreps. 1979. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20: 381–408.

Hart, O., and J. Moore. 1998. Default and renegotiation: A dynamic model of debt. *Quarterly Journal of Economics* 113: 1–41.

Harvey, C.R. 1989. Time-varying conditional covariances in tests of asset pricing models. *Journal of Financial Economics* 24: 289–317.

Hasbrouck, J. 1991. Measuring the information content of stock trades. *Journal of Finance* 46: 179–207.

He, H., and D.M. Modest. 1995. Market frictions and consumption-based capital asset pricing. *Journal of Political Economy* 103: 94–117.

He, H., and J. Wang. 1995. Differential information and dynamic behavior of stock trading volume. *Review of Financial Studies* 8: 919–972.

Heaton, J., and D. Lucas. 1996. Evaluating the effects of incomplete markets on risk sharing and asset pricing. *Journal of Political Economy* 104: 443–487.

Hellwig, M. 1980. On the aggregation of information in competitive markets. *Journal of Economic Theory* 22: 477–498.

Hennessy, C.A. 2004. Tobin's Q, debt overhang, and investment. *Journal of Finance* 59: 1717–1742.

Heston, S.L. 1993. A close form solution for options with stochastic volatility. *Review of Financial Studies* 6: 327–343.

Holmstrom, B. 1999. Managerial incentive problems: A dynamic perspective. *Review of Economic Studies* 66: 169–182.

Hong, H., and J.C. Stein. 1999. A unified theory of underreaction, momentum trading and overreaction in asset markets. *Journal of Finance* 54: 2143–2184.

Hong, H., and J. Wang. 2000. Trading and returns under periodic market closures. *Journal of Finance* 55: 297–354.

Hoshi, T., A. Kashyap, and D. Scharfstein. 1991. Corporate structure, liquidity, and investment: Evidence from Japanese industrial groups. *Quarterly Journal of Economics* 106: 33–60.

Hovakimian, A., T. Opler, and S. Titman. 2001. The debt–equity choice. *Journal of Financial and Quantitative Analysis* 36: 1–24.

Huang, J., and J. Wang. 1997. Market structure, security prices and informational efficiency. *Macroeconomic Dynamics* 1: 169–205.

Huang, J., and J. Wang. 2006a. Liquidity, asset prices and welfare under costly participation. Working paper. Massachusetts Institute of Technology.

Huang, J., and J. Wang. 2006b. Liquidity and market crashes. Working paper. Massachusetts Institute of Technology.

Hull, J., and A. White. 1987. The pricing of options on assets with stochastic volatility. *Journal of Finance* 42: 281–300.

Hull, J., and A. White. 1994. Numerical procedures for implementing term structure models II: Two-factor models. *Journal of Derivatives* 2: 37–48.

Ikenberry, D., J. Lakonishok, and T. Vermaelen. 1995. Market underreaction to open market share repurchases. *Journal of Financial Economics* 39: 181–208.

Jagannathan, R., and Z. Wang. 1996. The conditional CAPM and the cross-section of expected returns. *Journal of Finance* 51: 3–53.

Jarrow, R.A., and S.M. Turnbull. 1995. Pricing derivatives on financial securities subject to credit risk. *Journal of Finance* 50: 53–85.

Jegadeesh, N., and S. Titman. 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance* 48: 65–91.

Jensen, M.C. 1986. Agency costs of free cash flow, corporate finance, and takeovers. *American Economic Review* 76: 323–339.

Jensen, M.C., and W.H. Meckling. 1976. Theory of firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.

Johannes, M. 2004. The statistical and economic role of jumps in continuous-time interest rate models. *Journal of Finance* 59: 227–260.

Jones, C.M., and O.A. Lamont. 2002. Short sale constraints and stock returns. *Journal of Financial Economics* 66: 207–239.

Jones, C.P., and R.H. Litzenberger. 1970. Quarterly earnings reports and intermediate stockprice trends. *Journal of Finance* 25: 143–148.

Jung, K., Y. Kim, and R.M. Stulz. 1996. Timing, investment opportunities, managerial discretion, and the security issue decision. *Journal of Financial Economics* 42: 159–186.

Kahneman, D., and A. Tversky. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185: 1124–1131.

Kaplan, S.N., and P. Stromberg. 2001. Financial contracting theory meets the real world: An empirical analysis of venture capital contracts. *Review of Economic Studies* 70: 281–315.

Kennedy, D.P. 1994. The term structure of interest rates as a Gaussian random field. *Mathematical Finance* 4: 247–258.

Knight, F.H. 1936. The quantity of capital and the rate of interest. *Journal of Political Economy* 44(433–63): 612–642.

Kogan, L. 2001. An equilibrium model of irreversible investment. *Journal of Financial Economics* 62: 201–245.

Kogan, L., S.A. Ross, J. Wang, and M.M. Westerfield. 2006. The price impact and survival of irrational traders. *Journal of Finance* 61: 195–229.

Kothari, S.P., J. Shanken, and R.G. Sloan. 1995. Another look at the cross-section of expected returns. *Journal of Finance* 50: 185–224.

Krishnamurthy, A. 2002. The bond/old-bond spread. *Journal of Financial Economics* 66: 463–506.

Kyle, A.S. 1985. Continuous auctions and insider trading. *Econometrica* 53: 1315–1335.

Kyle, A.S., and W. Xiong. 2001. Contagion as a wealth effect. *Journal of Finance* 56: 1401–1440.

Lando, D. 1998. On Cox processes and credit-risky securities. *Review of Derivatives Research* 2: 99–120.

Lang, L., E. Ofek, and R.M. Stulz. 1996. Leverage, investment, and firm growth. *Journal of Financial Economics* 40: 3–29.

Langetieg, T.C. 1980. A multivariate model of the term structure of interest rates. *Journal of Finance* 35: 71–97.

Lee, C.M.C., A. Shleifer, and R.H. Thaler. 1991. Investor sentiment and the closed-end fund puzzle. *Journal of Finance* 46: 75–109.

Lehmann, B.N., and D.M. Modest. 1988. The empirical foundations of the arbitrage pricing theory. *Journal of Financial Economics* 21: 213–254.

Leland, H.E. 1994. Risky debt, bond covenants and optimal capital structure. *Journal of Finance* 49: 1213–1252.

Leland, H.E., and D.H. Pyle. 1977. Informational asymmetries, financial structure, and financial intermediation. *Journal of Finance* 32: 371–387.

Leland, H.E., and K.B. Toft. 1996. Optimal capital structure, endogenous bankruptcy and the term structure of credit spreads. *Journal of Finance* 50: 789–819.

LeRoy, S.F., and R.D. Porter. 1981. The present-value relation: Tests based on implied variance bounds. *Econometrica* 49: 555–574.

Lesmond, D.A., M.J. Schill, and C. Zhou. 2004. The illusory nature of momentum profits. *Journal of Financial Economics* 71: 349–380.

Lettau, M., and S. Ludvigson. 2001. Resurrecting the (C) CAPM: A cross-sectional test when risk premia are time-varying. *Journal of Political Economy* 109: 1238–1287.

Lewellen, J., and S. Nagel. 2006. The conditional CAPM does not explain asset-pricing anomalies. *Journal of Financial Economics* 82: 289–314.

Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics* 47: 13–37.

Litterman, R., and J. Scheinkman. 1991. Common factors affecting bond returns. *The Journal of Fixed Income* 1: 54–61.

Liu, J., J. Pan, and T. Wang. 2005. An equilibrium model of rare-event premia and its implication for option smirks. *Review of Financial Studies* 18: 131–164.

Llorente, G., R. Michaely, G. Saar, and J. Wang. 2002. Dynamic volume–return relation of individual stocks. *Review of Financial Studies* 15: 1005–1047.

Lo, A.W., and A.C. Mackinlay. 1988. Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies* 1: 41–66.

Lo, A.W., and A.C. MacKinlay. 1990. When are contrarian profits due to stock market overreaction? *Review of Financial Studies* 3: 175–205.

Lo, A.W., H. Mamaysky, and J. Wang. 2004. Asset prices and trading volume under fixed transactions costs. *Journal of Political Economy* 112: 1054–1090.

Lo, A.W., and J. Wang. 2006. Trading volume: Implications of an intertemporal capital asset pricing model. *Journal of Finance* 61: 2805–2840.

Longstaff, F.A. 1989. A nonlinear general equilibrium model of the term structure of interest rates. *Journal of Financial Economics* 23: 195–224.

Longstaff, F.A., and E.S. Schwartz. 1992. Interest rate volatility and the term structure: A two factor general equilibrium model. *Journal of Finance* 47: 1259–1282.

Longstaff, F.A., and E.S. Schwartz. 1995. A simple approach to valuing risky fixed and floating rate debt. *Journal of Finance* 50: 789–819.

Lucas, R.E. Jr. 1978. Asset prices in an exchange economy. *Econometrica* 46: 1429–1445.
Luttmer, E.G.J. 1999. What level of fixed costs can reconcile consumption and stock returns? *Journal of Political Economy* 107: 969–997.
Lyons, R.K. 1995. Tests of microstructural hypotheses in the foreign exchange market. *Journal of Financial Economics* 39: 321–351.
Madhavan, A., and S. Smidt. 1993. An analysis of changes in specialist inventories and quotations. *Journal of Finance* 48: 1595–1628.
Maksimovic, V., and S. Titman. 1991. Financial policy and reputation for product quality. *Review of Financial Studies* 4: 175–200.
Malkiel, B.G. 1977. The valuation of closed-end investment company shares. *Journal of Finance* 32: 847–859.
Markowitz, H.M. 1952. Portfolio selection. *Journal of Finance* 7: 77–91.
McConnell, J.J., and C.J. Muscarella. 1985. Corporate capital expenditure decisions and the market value of the firm. *Journal of Financial Economics* 14: 399–422.
Mehra, R., and E.C. Prescott. 1985. The equity premium: A puzzle. *Journal of Monetary Economics* 15: 145–161.
Melino, A., and S.M. Turnbull. 1990. Pricing foreign currency options with stochastic volatility. *Journal of Econometrics* 45: 239–265.
Merton, R.C. 1973. The theory of rational option pricing. *Bell Journal of Economics and Management Science* 4: 141–183.
Merton, R.C. 1974. On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29: 449–470.
Merton, R.C. 1976. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3: 125–144.
Merton, R.C. 1981. On market timing and investment performance part I: An equilibrium theory of value for market forecasts. *Journal of Business* 54: 363–406.
Merton, R.C. 1987. A simple model of capital market equilibrium with incomplete information. *Journal of Finance* 42: 483–510.
Milgrom, P., and N. Stokey. 1982. Information, trade and common knowledge. *Journal of Economic Theory* 26: 17–27.
Miller, E.M. 1977a. Risk, uncertainty, and divergence of opinion. *Journal of Finance* 32: 1151–1168.
Miller, M.H. 1977b. Debt and taxes. *Journal of Finance* 32: 261–276.
Mitchell, M., L.H. Pedersen, and T. Pulvino. 2007. Slow moving capital. *American Economic Review* 97: 215–220.
Modigliani, F., and M.H. Miller. 1958. The cost of capital, corporate finance and the theory of investment. *American Economic Review* 48: 261–297.
Modigliani, F., and M.H. Miller. 1963. Corporate income taxes and the cost of capital: A correction. *American Economic Review* 53: 433–443.
Myers, S.C. 1977. Determinants of corporate borrowing. *Journal of Financial Economics* 5: 147–176.
Myers, S.C. 2000. Outside equity. *Journal of Finance* 55: 1005–1037.
Myers, S.C., and N.S. Majluf. 1984. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of Financial Economics* 13: 187–221.
Narayanan, M.P. 1985. Managerial incentives for short-term results. *Journal of Finance* 40: 1469–1484.
O'Hara, M. 2003. Presidential address: Liquidity and price discovery. *Journal of Finance* 58: 1335–1354.
Odean, T. 1998. Are investors reluctant to realize their losses? *Journal of Finance* 53: 1775–1798.
Pan, J. 2002. The jump-risk premia implicit in options: Evidence from an integrated time-series study. *Journal of Financial Economics* 63: 3–50.
Pastor, L., and R. Stambaugh. 2003. Liquidity risk and expected stock returns. *Journal of Political Economy* 113: 642–685.
Petkova, R., and L. Zhang. 2005. Is value riskier than growth? *Journal of Financial Economics* 78: 187–202.
Phillips, G.M. 1995. Increased debt and industry product markets: An empirical analysis. *Journal of Financial Economics* 37: 189–238.
Piazzesi, M. 2005. Bond yields and the Federal Reserve. *Journal of Political Economy* 113: 311–344.
Poterba, J., and L.J. Summers. 1988. Mean reversion in stock prices: Evidence and implications. *Journal of Financial Economics* 22: 27–59.
Rajan, R., and L. Zingales. 1995. What do we know about capital structure? Some evidence from international data. *Journal of Finance* 50: 1421–1460.
Ready, M.J. 1999. The specialist's discretion: Stopped orders and price improvement. *Review of Financial Studies* 12: 1075–1112.
Rietz, T.A. 1988. The equity risk premium: A solution. *Journal of Monetary Economics* 21: 117–131.
Rock, K. 1990. The specialist's order book and price anomalies. Working paper. Harvard Business School.
Roll, R. 1977. A critique of the asset pricing theory tests. Part I: On past and potential testability of the theory. *Journal of Financial Economics* 4: 129–176.
Roll, R. 1984a. Orange juice and weather. *American Economic Review* 74: 861–880.
Roll, R. 1984b. A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance* 39: 1127–1139.
Roll, R. 1988. R^2. *Journal of Finance* 43: 541–566.
Romer, D. 1993. Rational asset-price movements without news. *American Economic Review* 83: 1112–1130.
Ross, S.A. 1973. The economic theory of agency: The principals problems. *American Economic Review* 63: 134–139.
Ross, S.A. 1976. Options and efficiency. *Quarterly Journal of Economics* 90: 75–89.
Ross, S.A. 1977. The determination of financial structure: The incentive signaling approach. *Bell Journal of Economics* 8: 23–40.

Ross, S.A. 2002. Neoclassical finance, alternative finance, and the closed end fund puzzle. *European Financial Management* 8: 129–137.

Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics* 7: 407–425.

Rubinstein, M. 1994. Implied binomial tress. *Journal of Finance* 49: 771–818.

Saa-Requejo, J., and P. Santa-Clara. 1999. Bond pricing with default risk. Working paper. UCLA.

Samuelson, P.A. 1965. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6: 41–49.

Sandroni, A. 2000. Do markets favor agents able to make accurate predictions. *Econometrica* 68: 1303–1341.

Santa-Clara, P., and D. Sornette. 2001. The dynamics of the forward interest rate curve with stochastic string shocks. *Review of Financial Studies* 14: 149–185.

Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.

Schaefer, S.M., and E.S. Schwartz. 1984. A two-factor model of the term structure: An approximate analytical solution. *Journal of Financial and Quantitative Analysis* 19: 413–424.

Scheinkman, J.A., and W. Xiong. 2003. Overconfidence and speculative bubbles. *Journal of Political Economy* 111: 1183–1219.

Scott, L.O. 1997. Pricing stock options in a jump-diffusion model with stochastic volatility and interest rates: Applications of Fourier inversion methods. *Mathematical Finance* 7: 413–426.

Seppi, D.J. 1997. Liquidity provision with limit orders and a strategic specialist. *Review of Financial Studies* 10: 103–150.

Shanken, J. 1982. The arbitrage pricing theory? It is testable. *Journal of Finance* 37: 1129–1140.

Shanken, J. 1987. Multivariate proxies and asset pricing relations: Living with the Roll critique. *Journal of Financial Economics* 18: 91–110.

Shanken, J. 1990. Intertemporal asset pricing: An empirical investigation. *Journal of Econometrics* 45: 99–120.

Sharpe, W.F. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.

Shiller, R.J. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* 71: 421–436.

Shleifer, A., and R.W. Vishny. 1997. The limits of arbitrage. *Journal of Finance* 52: 35–55.

Stambaugh, R.F. 1982. On the exclusion of assets from tests of the two parameter model. *Journal of Financial Economics* 10: 235–268.

Stanton, R. 1997. A nonparametric model of term structure dynamics and the market price of interest rate risk. *Journal of Finance* 52: 1973–2002.

Stein, E.M., and C.J. Stein. 1991. Stock price distributions with stochastic volatility: An analytic approach. *Review of Financial Studies* 4: 727–752.

Stoll, H.R. 1979. *Regulation of securities markets: An examination of the effects of increased competition*. New York: Graduate School of Business/New York University.

Stulz, R. 1990. Managerial discretion and optimal financing policies. *Journal of Financial Economics* 26: 3–27.

Sundaresan, S.M. 1989. Intertemporally dependent preferences and the volatility of consumption and wealth. *Review of Financial Studies* 2: 73–88.

Taggart, R.A. 1977. A model of corporate financing decision. *Journal of Finance* 32: 1467–1484.

Titman, S., and R. Wessels. 1988. The determinants of capital structure choice. *Journal of Finance* 43: 1–19.

Tobin, J. 1958. Liquidity preference as behavior towards risk. *Review of Economic Studies* 25: 65–86.

Townsend, R.M. 1978. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 21: 265–293.

Tuckman, B., and J.L. Vila. 1992. Arbitrage with holding costs: A utility based approach. *Journal of Finance* 47: 1283–1302.

Vasicek, O. 1977. An equilibrium characterization of the term structure. *Journal of Financial Economics* 5: 177–188.

Vayanos, D. 1998. Transaction costs and asset prices: A dynamic equilibrium model. *Review of Financial Studies* 11: 1–58.

Vayanos, D. 1999. Strategic trading and welfare in a dynamic market. *Review of Economic Studies* 66: 219–254.

Vayanos, D., and P. O. Weill. 2006. A search-based theory of the on-the-run phenomenon. Working Paper No. 12670. Washington, DC: NBER.

von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. New York: Wiley.

Wahal, S. 1997. Entry, exit, market makers and the bid-ask spread. *Review of Financial Studies* 10: 871–901.

Wald, J. 1999. Capital structure with dividend restrictions. *Journal of Corporate Finance* 5: 193–208.

Wang, J. 1993. A model of intertemporal asset prices under asymmetric information. *Review of Economic Studies* 60: 249–282.

Wang, J. 1994. A model of competitive stock trading volume. *Journal of Political Economy* 102: 127–168.

Wang, J. 1996. The term structure of interest rates in a pure exchange economy with heterogeneous investors. *Journal of Financial Economics* 41: 75–110.

Wang, K.Q. 2003. Asset pricing with conditioning information: A new test. *Journal of Finance* 58: 161–196.

Weil, P. 1989. The equity premium puzzle and the risk-free rate puzzle. *Journal of Monetary Economics* 24: 401–421.

Weinstein, N.D. 1980. Unrealistic optimism about future life events. *Journal of Personality and Social Psychology* 39: 806–820.

Weiss, L.A., and K.H. Wruck. 1998. Information problems, conflicts of interest and asset stripping: Chapter 11's failure in the case of eastern airlines. *Journal of Financial Economics* 48: 55–97.

Williamson, O.E. 1964. *The economics of discretionary behavior: Managerial objectives in a theory of the firm*. Englewood Cliffs: Prentice-Hall.
Williamson, O.E. 1975. *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press.
Wilson, R. 1968. On the theory of syndicates. *Econometrica* 36: 119–132.
Yuan, K. 2005. Asymmetric price movements and borrowing constraints: A rational expectations equilibrium model of crises, contagion, and confusion. *Journal of Finance* 60: 379–411.
Zhang, L. 2006. Anomalies. Working Paper No. 11322. Cambridge, MA: NBER.
Zingales, L. 1998. Survival of the fittest or the fattest? Exit and financing in the trucking industry. *Journal of Finance* 53: 905–938.
Zwiebel, J. 1996. Dynamic capital structure under managerial entrenchment. *American Economic Review* 86: 1197–1215.

Finance and Saving

Victoria Chick

Saving and finance are now clearly distinguished (though perhaps surprisingly this is a fairly recent development). Finance refers to monetary transactions securing the means of payment for purchases in excess of current cash flow or funding the holding of assets. Problems of finance exist for individuals or firms, not for economies as a whole except in relation to other countries; accordingly, the analysis of finance is microeconomic in character.

Saving is income not consumed. In contrast to finance it is both an action undertaken by individuals and an outcome for the economy as a whole. In the context of today's financial institutions, individual saving largely consists of money-flows to those institutions, though saving can take 'real' form as well (e.g. the purchase of houses or works of art). For an economy as a whole, apart from its interactions with the rest of the world, saving can only be net capital accumulation. Financial transactions 'consolidate out' in the aggregate balance sheet.

The problem for theory is to bridge the gap between microeconomic money-flows, arising out of individual acts of saving and constituting potential provision of finance, and the amount of actual saving at the macroeconomic level. (Even in money terms the sum of individual attempts to save will not necessarily 'add up' to aggregate saving, for revaluations which occur as savers compete for assets are hidden in the aggregate.)

There was a time when this gap between the intentions of individuals and the macroeconomic outcome was not perceived. In Ricardo's model (1817) of an economy producing one staple good (corn), the corn not consumed was seed-corn, saving and investment in one, with no need for finance. The amount of corn so saved would depend on the expected rate of return from forgoing consumption.

Ricardo's formulation, not inappropriate to a largely non-monetized ('real') agricultural economy, survives to this day in the form of time-preference theory, in which consumption and saving are seen as two sides of an intertemporal consumption plan. The trade-off between the disutility of deferred consumption and the expected rate of return on investment determines the volume of saving and investment. Monetary factors, borrowing and lending, can be added to this theory if the rate of interest is distinguished from the rate of profit. The rate of interest in this analysis is taken as exogeneous since the theory pertains to individual choice.

Amongst the classical economists with the exception of Marx, and indeed in much current theory, the distinction between interest and profit is imperfectly made. Marx (1867) was much concerned with the conditions under which finance capital could be obtained and sufficient profit on industrial capital realized to pay back finance capital, a problem also central to the work of Keynes.

The immediate background into which Keynes's work was inserted consisted of Wicksell's theory and the loanable funds theory associated with Wicksell's successors in Stockholm and with Keynes's colleague, Sir Dennis Robertson. These theories added monetary factors to the classical theory, in which the rate of interest/profit was determined by the equality of 'real' saving and investment.

Wicksell (1901) proposed the concept of the 'natural rate of interest', the rate compatible with

saving–investment equality, and contrasted this with the money rate of interest. A divergence of the natural rate from the money rate of interest would result in a cumulative process of inflation or deflation caused by expansion or contraction of bank credit. The process would converge as the two rates once again became equal. Hence equality of the natural and money rates yielded price stability.

Unfortunately the concept of the natural rate is not observable nor is it determinate independently of the level of employment. It is now regarded as unhelpful.

In loanable funds theory the money rate of interest is determined by equating the demand for loanable funds, defined to include 'hoarding' (additions to idle money-holdings) as well as investment, with the supply of funds, comprising saving and additions to the money supply. In this theory as in Wicksell's, saving is implicitly equivalent to providing finance. Hoarding, clearly not a source of finance, is also not a form of saving. Saving, however, is not the only source of finance, there is also bank credit. Increases in the money supply occur when banks expand credit by more than 'prior saving' in the form of deposits. These increases were, as in Wicksell, generally held to be inflationary.

From a microeconomic perspective it would seem plausible that saving, money-income not consumed and thus available for lending to deficit spenders, automatically constitutes finance. To Keynes (1936), however, hoarding was a form of saving, and hoarding does not provide finance. Also, it is implicit in his theory of speculative demand, in which the determination of the rate of interest is dominated by trade in existing securities, that saving does not provide finance if the pace of new issues is not adequate to absorb the savings flow. Thus for the first time the theory of savings was divorced from the theory of finance.

Until Keynes, investment was assumed to be dependent on saving as the source of finance. Keynes reversed this causal ordering, arguing that investment, financed independently of saving, created additional income adequate eventually to generate an equal volume of saving. Robertson (1940) demonstrated that the source of finance must be bank lending. This results in an increase of deposits the holding of which, on Keynes's definition but not in Robertson's, constituted saving. Much of the long debate between these two clever economists (see Keynes 1973, pp. 201–34) rested on a misunderstanding.

Today the theory of saving has developed little beyond the debates on the relative determining roles of rates of interest and levels of income which dominated the subject in the 1930s. By contrast, the theory of finance, dealing with the appropriate portfolio choices of active managers of financial portfolios and the options open to firms in the finance of their capital, has been much developed and refined and is full of vitality.

See Also

- ▶ Keynes, John Maynard (1883–1946)
- ▶ Loanable Funds
- ▶ Saving Equals Investment

Bibliography

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

Keynes, J.M. 1973. *The collected writings of John Maynard Keynes*, vol. XIV. London: Macmillan.

Marx, K. 1867. *Capital*. Hamburg: Otto Meisner.

Ricardo, D. 1817. *Principles of political economy and taxation*,, 1971. Harmondsworth: Penguin Books.

Robertson, D.H. 1940. Effective demand and the multiplier. In *Essays in monetary theory*, ed. D.H. Robertson. London: P.S. King.

Wicksell, K. 1901. *Lectures in political economy*,, 1934. London: Routledge & Kegan Paul.

Finance Capital

J. Tomlinson

The concept of finance capital encapsulates the most theoretically significant attempt by the orthodox Marxism of the pre-1914 period to come to terms with the developments of

capitalism in the late 19th century. After the Bolshevik Revolution the concept was much less frequently employed. In part this demise reflected the breakdown of orthodox Marxism as a relatively unified but developing body of doctrine, but it also reflected the inherent problems of the concept.

The term itself is not to be found in Marx's work. But subsequent formulations relied heavily on the schematic outline by Marx in Part V of Volume III of *Capital*, especially chapter 27 on 'The Role of Credit in Capitalist Production'. Marx's arguments, penned in the 1860s, but not published until 1894, focus on the two processes of the multiplication of forms of credit available to industrial capital, and the formation of joint stock companies. The two processes together he saw as heralding 'the abolition of capital as private property within the framework of capitalist production itself ' (1894, p. 436).

On the basis of Marx's brief outline, Hilferding in his *Finanz Kapital* (1908), built a systematic argument, conceiving finance capital as the highest stage of capitalism. Hilferding's book presents a theoretical history of the evolution of relations between money and productive (industrial) capital. This relationship is seen as having gone through a series of historical transformations, particularly on the basis of changes in the form of credit and credit-giving institutions. Trade credit (or 'circulation credit') is seen as the initial form of credit, emerging from interruptions to the cycle of capital, and tying credit creation directly to the production and sale of commodities. This form of credit facilitated an extension of the scale of production by using funds otherwise idle.

Subsequently there developed banks which not only recycled capitalists' own idle funds but put money from other sources at the disposal of industrial capitalists. When this process of credit expansion encompassed the financing of fixed capital the relationship of the banks to industrial capital began to change, as banks came to have an enduring rather than a momentary interest in the fortunes of the industrial enterprise they lent to. So emerged the characteristically 'German' interlinking of banks and industry, with banks controlling large blocks of industrial equity and sharing large numbers of directors with industry.

The changing relationship encouraged the growth of larger banks, which could afford to tie up funds in this way, but also were enabled by expansion in size to finance lots of firms in order to spread their risks. This growing concentration of banks was seen as interacting with the growth of concentration amongst industrial firms, and is thereby closely linked with the development of the joint stock company. The growth of shares, which Hilferding stresses should be seen as another form of (irredeemable) credit, is a pre-condition of the growth of the joint stock company, which in turn is a pre-condition of a full utilization of the possibilities of technological advance (pp. 122–3).

These joint stock companies become more and more concentrated and tend to the elimination of free competition. This is paralleled by the growth of 'an ever more intimate relationship' between banks and industrial capital: 'Through this relationship . . . capital assumes the form of finance capital, its supreme and most abstract expression' (p. 21).

But for Hilferding finance capital is not just a concept but a real social and political force (as indeed it was in Germany). It has its own economic policies, which are both protective of the home market and promote expansion abroad. This latter impetus leads to an intimate relationship between finance capital and the state, which is used to pursue policies of territorial aggrandizement, built partly on the desire to export commodities, but above all to facilitate the export of capital. Hence the characteristic ideology of finance capital (unlike competitive industrial capital) is aggressively expansionist and aspires to political as well as economic domination. 'Thus the ideology of imperialism arises on the ruins of the old liberal ideals, whose naivety it derides' (p. 334).

Hilferding's analysis of the structure of finance capital can be read as largely a Marxist version of the well known story of the 'divorce of ownership and control' via the development of the joint stock company. Such a parallel would not be entirely misplaced, but it would obscure some of the most important elements of Hilferding's theories.

Least surprisingly, Hilferding's analysis deploys Marx's theory of value, and this, for example, leads him to picture finance capital seizing profits originally produced by industrial capital. Such analysis simply reflects the Marxist concept of industrial capital as productive of surplus value, with other capitals obtaining their profits by redistribution from this original source. But the conceptual background of Marx's theory of value has more specific implications for Hilferding's work.

The argument that values and profits arise originally only in the industrial sector leads to the characterization of share capital as 'fictitious' capital (a term also deployed by Marx), compared with 'genuinely functioning industrial capital' (p. 111). This essentially moralistic approach cuts across the useful discussion by Hilferding of the role of share capital in making possible the joint stock company form of organization, with the progressiveness of this form for the development of production. Similarly, this allegiance to the primacy of industrial capital leads him to assert that 'the techniques of banking itself generate tendencies which affect the concentration of the banks and industry alike, but the concentration of industry is the ultimate cause of concentration in the banking system' (p. 98). Yet his analysis elsewhere makes clear that the development of the banking system, and credit system more generally, were more commonly pre-conditions of the development of forms of industrial capital than vice versa.

A problem of a rather different order is Hilferding's treatment of the relationship between banks and industry as the defining characteristic of finance capital. This leads to the view that countries such as England, where these close relations never existed, are deviants from the norm of development: '. . . the English system is an outmoded one and is everywhere on the decline because it makes control of the loaned-out bank capital more difficult, and hence obstructs the expansion of bank capital itself' (p. 293). But Hilferding's own arguments on the stock exchange, as the basis of a particular form of credit creation, undercuts this identification of finance capital with one particular financial institution – banks. For what is clearly at stake in Hilferding's general arguments is the development of different types of *credit*, which then impinges on forms of industrial organization, but where these types are not tied to any particular institutional form. (This is quite clear in most of his discussion of the stock exchange.)

Hilferding thus imparts a strong evolutionary element into his argument, where the normal path of development is towards the 'German' model of the relationship between banks and industry. This evolutionism is also more broadly present in Hilferding when he follows Marx in seeing the growth of finance capital and the joint stock company as a socialization of production, that is, a step towards socialist organization of the economy. This socialization is theorized as consisting of a development of a complex division of labour organized by a very few sites of decision-making. Hence the struggle for socialism in this framework is reduced to a struggle to dispossess the oligarchy who currently control production, but who have unwittingly created the 'final organizational prerequisites for socialism' (p. 368). This extraordinary line of argument implies that there is nothing specifically capitalist about the organization of large-scale capitalist industry, except who controls it – surely a *reductio ad absurdum* of the notion of the productive forces developing independently of the relations of production.

The concept of finance capital was most famously deployed by Lenin in his work on Imperialism. Lenin's aim was quite clearly to engage in a political polemic not a theoretical analysis, and he adds nothing new to the discussion of the concept. His main difference with Hilferding was to take further the stress on the aggressive tendencies of finance capital, and to argue the inescapability of imperialist war in such conditions, a conclusion not drawn by Hilferding. Whatever the merits or otherwise of Lenin's political polemic, the association of Hilferding's work with it tended to obscure the theoretical significance of *Finanz Kapital*.

After Lenin, the concept of Finance Capital has played a much lesser role in Marxist discussions. Instead, Bolshevized Marxism has tended to place more emphasis on the monopoly characteristics of

modern capitalism, rather than the finance aspect; hence the common deployment of concepts of Monopoly Capital, and State Monopoly Capitalism. But even within the conceptual approaches of this post-1917 orthodox Marxism this emphasis appears misplaced. As Hussain (1976) has convincingly argued, in terms of standard Marxist categories the concept of finance capital provides a basis for the periodization of capitalism which monopoly capital cannot. It is the relationship of finance of industrial capital which largely determines the structure and size of firms, and hence finance determines the level of 'monopoly'. Starting with the total social capital, as Marx does, it is the relation of finance to industrial capital which determines the distribution of capital into firms. Within an orthodox Marxist framework, finance could in this way provide a basis for periodizing capitalism, that is, on the basis of changes in the relationship of finance to industrial capital and their implications.

Hilferding's work shares some of the defects of Marx's *Capital* in which it was so clearly grounded. Its evolutionism and its adherence to Marx's theory of value, in particular, tend to obscure what is most valuable in the analysis. Nevertheless, with the growing prominence of financial institutions and financial calculation in advanced capitalist countries, any work which provides a detailed theoretical study of the workings of finance under capitalism needs to be taken seriously. This is especially so when the study, at its best, provides analyses which avoid both the speculative character of discussion of the 'total social capital', and the empiricism of institutional description. Rather, the concept of finance capital provides an entry into analysing the nexus of relationships between financial and industrial institutions, but where these institutions are seen neither as simply representations of broader social forces, nor as complex entities knowable only through description.

More specifically, the concept of finance capital leads us to treat the industrial structure as an effect of the changes in the relationship between industrial and financial capital. Thus, for example, the well-known growth of industrial concentration in the UK and other countries in the 1950s and 1960s would be analysed primarily as an effect of the operations of the stock market, and of the credit-creating criteria deployed in that market. Equally, prediction of future trends in the industrial structure would depend upon views about the future evolution of the financial system. The development of the industrial structure, seen in this light, would neither be technologically determined, as commonly suggested, nor, as in some Marxist treatments, would it be seen as tied to the idea of the appropriation by a new class of capitalist of power over the means of production. Rather, the focus would be on the conditions of existence of the credit-giving criteria employed by financial institutions, and how these structured the forms of calculation used by firms in their deployment of means of production. In this way, forms of calculation would be seen as central to the analysis of capitalist firms, but where these forms were themselves seen as dependent upon the mechanisms of allocation of credit in the economy.

It would be an impossible project to 'revive' the orthodox Marxism of the pre-1914 period. Its theoretical presuppositions are in crucial respects no longer tenable, and its specific analyses often tied to circumstances which have changed out of all recognition. Nevertheless, this was a period when Marxism was a relatively open programme of research, and the results of that are not to be simply discarded. The concept of finance capital, shorn of some of its theoretical baggage, could be seen as a potentially fruitful legacy from that period.

See Also

- ▶ Hilferding, Rudolf (1877–1941)
- ▶ Monopoly Capitalism

Bibliography

Hilferding, R. 1908. *Finanz Kapital*. Translated into English, with an introduction by T.B. Bottomore. London: Routledge & Kegan Paul, 1981.

Hussain, A. 1976. Hilferding's Finance Capital. *Bulletin of the Conference of Socialist Economists*.

Marx, K. 1894. *Capital*, vol. III. Ed. F. Engels. London: Lawrence & Wishart, 1968.

Financial Accelerator

Oliver de Groot

Abstract

The financial accelerator refers to the mechanism by which distortions (frictions) in financial markets amplify the propagation of shocks through an economy. This article sets out the theoretical foundations of the financial accelerator in financial friction DSGE (Dynamic Stochastic General Equilibrium) models and discusses the ability of these models to provide policy recommendations and a narrative for the 2007–08 financial crisis.

Keywords

Business Cycles; Calibration; Dynamic Macroeconomics; Dynamic Stochastic General Equilibrium (DSGE) Models; Estimation; Expectations; Identification; Intertemporal Optimisation Problems; Monetary Policy Shocks; Technology Shocks; Linear Models

JEL Classifications

G63; E32; E44; E52; G11

Introduction

The financial accelerator refers to the mechanism by which frictions in financial markets amplify the propagation of shocks through an economy. With the financial accelerator, an initial deterioration in credit market conditions leads to rising credit spreads, creating an additional weakening of credit market conditions and resulting in a disproportionately large drop in economic activity.

The key building block of the financial accelerator is the existence of a friction in the intermediation of credit. In frictionless financial markets, loanable funds are intermediated efficiently between savers and borrowers. Furthermore, in line with the insights of Modigliani and Miller (1958), the composition of borrowers' internal (own net worth) and external (borrowed) funds does not affect real economic outcomes. However, in reality, asymmetric information and imperfect contract enforcement creates principal–agent problems between borrowers and lenders. The Modigliani–Miller theorem no longer holds and fluctuations in borrower net worth have real economic consequences. The mechanism involves an inverse relationship between credit spreads (the cost of borrowing over the risk-free rate) and net worth. This inverse relationship arises because, when a borrower's net worth is low, the borrower's incentive to (for example) truthfully report returns, exert high effort or not abscond with assets is also low. As borrowers' and lenders' interests become more divergent, agency costs and hence credit spreads increase. To the extent that borrowers' net worth is procyclical, credit spreads will be countercyclical, with borrowing costs increasing in downturns, amplifying fluctuations in investment and economic activity.

While the term was first coined by Bernanke et al. (1996), the idea that credit market conditions play a central role in economic fluctuations has much earlier origins. Many economists who lived through the 1930s, including Fisher (1933), Keynes (1936), Kindleberger (1978) and Minsky (1992), believed that the financial sector – in excessively curtailing lending in response to falling asset prices – was largely responsible for the Great Depression.

By the 1980s, *real business cycle* models, with frictionless financial markets, dominated the macroeconomic research agenda. However, large fluctuations in economic activity often appeared to result from small disturbances and real business cycle models struggled to generate the propagation and amplification necessary to match this observation. The financial accelerator mechanism – by introducing a distortion in the credit market of an otherwise standard real business cycle model – was one solution to this 'small disturbances, large fluctuations' puzzle.

The original microfoundation of the financial accelerator, in Bernanke and Gertler (1989) (and popularised by the quantitative business cycle

framework of Bernanke et al. (1999)), was based on the 'costly state verification' problem of Townsend (1979), in which costly bankruptcy resulted from an asymmetry of information between lenders and borrowers. A number of alternative microfoundations have since emerged. Kiyotaki and Moore (1997) generated credit cycles when lenders faced the 'hold-up' problem studied by Hart and Moore (1994), giving rise to collateral constraints on borrowing. Both of these early contributions to the financial accelerator literature focused on non-financial borrowers. Since the 2007–08 financial crisis, however, many models have focused instead on the problems faced by financial intermediaries (banks) in obtaining funds. Most popular among them, Gertler and Karadi (2011) proposed the so-called 'running away' moral hazard problem, in which bankers' ability to abscond with assets endogenously limits bank leverage.

In addition to these, Christiano and Ikeda (2013) introduced a microfounded financial accelerator by adopting an unobserved effort moral hazard problem on the part of borrowers, de Groot (2010) used a (global games) coordination game between lenders in the spirit of Goldstein and Pauzner (2005), and Adrian and Shin (2014) introduced a Value-at-Risk constraint.

Despite the financial accelerator literature having become well established by the mid-2000s, Vlcek and Roger (2012) showed that financial frictions were almost non-existent in the quantitative DSGE models used by central banks and policy institutions at that time. The financial crisis naturally brought a renewed interest in adding these frictions to policy models to improve forecasting and provide insights for the design of monetary and macroprudential policy. However, while alternative microfoundations produce the same basic financial accelerator mechanism – in which deteriorating balance sheet conditions of borrowers exacerbate the agency problem, driving up credit spreads and depressing economic activity – each has advantages and disadvantages in terms of tractability and realism and no consensus approach has emerged. Identifying empirically the key friction in credit markets remains an important aspect of the research agenda.

The growth in the macro-finance literature since the financial crisis has been so large that this short survey cannot hope to do it all justice. This article will focus on a subset of the literature with models relying on linear approximation and frictions that always bind. Quadrini (2011), Christiano and Ikeda (2012) and Brunnermeier et al. (2013) survey the theoretical work on other financial instability phenomena, including occasionally binding constraints, fire sales, bank runs and pecuniary externalities.

The rest of this article will proceed as follows. The next section sketches a simple model of the financial accelerator without reference to a particular microfoundation. The subsequent section describes three prominent microfoundations. The final section asks: (1) How well do financial accelerator models fit the narrative of the 2007–08 financial crisis? (2) What are the policy implications of the financial accelerator? And (3) What challenges remain?

F

A Simple Model with a Financial Accelerator

In order to expose the heart of the financial accelerator mechanism – countercyclical credit spreads driven by procyclical fluctuations in borrowers' net worth – consider first the simplest DSGE model, a frictionless real business cycle model *à la* Brock and Mirman (1972). This model reduces to a single equilibrium condition for the loanable funds market.

There exists an infinitely lived representative household with log utility over consumption, $\mathbb{E}_t \sum_{t=0}^{\infty} \beta^t \log(c_t)$, where $\mathbb{E}_t(\cdot)$ is the expectations operator conditional on time t information, $\beta \in (0, 1)$ is the subjective discount factor and c_t is consumption. There also exists a representative firm with production technology, $y_t = \varepsilon_t k_{t-1}^{\alpha}$, where y_t is output, ε_t is a technology shock, k_{t-1} is the capital stock created in $t-1$ and productive in t, and $\alpha \in (0, 1)$. Household labour supply is fixed (and normalised to one) with real wages equal to the marginal product of labour. Capital fully depreciates each period, so market clearing is given by $\varepsilon_t k_{t-1}^{\alpha} = c_t + k_t$.

Suppose, for the purposes of story telling, there exists a competitive bank (ultimately owned by the household) intermediating loanable funds in a frictionless credit market in this economy. The household, as the supplier of loanable funds, saves via deposits and earns the gross risk-free return r_{t-1} at time t. The firm, as the demander of loanable funds for purchasing capital, borrows from the bank and pays the gross realised return on capital, $r_t^k = \alpha\varepsilon_t k_{t-1}^{\alpha-1}$. The (upward sloping) supply curve for loanable funds is sketched by the household's Euler equation, $1 = \mathbb{E}_t\beta(c_t/c_{t+1})r_t$, while the (downward sloping) demand curve is sketched by the expected marginal product of capital, $\mathbb{E}_t r_{t+1}^k = \alpha k_t^{\alpha-1}$.

Since the loanable funds market is frictionless, the bank is just a veil and the competitive equilibrium is the same as when households directly rent capital to firms. Arbitrage ensures that the expected discounted return on capital is equal to the discounted return on risk-free deposits,

$$\mathbb{E}_t\beta(c_t/c_{t+1})r_{t+1}^k = \mathbb{E}_t\beta(c_t/c_{t+1})r_t.$$

To a log-linear approximation there is no credit spread since the no-arbitrage condition becomes $\mathbb{E}_t\tilde{r}_{t+1}^k - \tilde{r}_t = 0$, where $\tilde{r}_t$, for example, denotes the log-linear deviation of r_t from steady state.

Consider next the response of this frictionless economy to a negative technology shock. On impact, the demand curve for loanable funds does not shift while the supply curve shifts inwards. To see this, substitute the no-arbitrage condition and the aggregate resource constraint into the Euler equation and derive the log-linear approximation of the supply curve

$$\mathbb{E}_t\tilde{r}_{t+1}^k = \underset{(-)(-)}{a\left(\tilde{\varepsilon}_t, \mathbb{E}_t\tilde{k}_{t+1}\right)} + b\tilde{k}_t,$$

where the intercept, a, is a decreasing function of $\tilde{\varepsilon}_t$ and $\mathbb{E}_t\tilde{k}_{t+1}$, and $b > 0$ is a positive slope coefficient (with both a and b functions of structural parameters). The negative shock, all else equal, reduces output (and consumption) at time t relative to $t + 1$, reducing the stochastic discount factor and therefore reducing the supply of loanable funds for any given expected return on capital. In equilibrium, the expected return on capital, $\mathbb{E}_t\tilde{r}_{t+1}^k$, rises and capital expenditure, $\tilde{k}_t$ falls.

How can we amplify the effect of the negative shock? Suppose there is – for some reason – a wedge (a credit spread) between the expected return on capital and the risk-free rate, $\tilde{s}_t \equiv \mathbb{E}_t\tilde{r}_{t+1}^k - \tilde{r}_t$, that is countercyclical. In other words, $\tilde{s}_t = \underset{(-)}{s\left(\tilde{\varepsilon}_t\right)}$ and, on impact of a negative shock, $\tilde{s}_t$ becomes positive. The supply curve for loanable funds using this limit-to-arbitrage condition becomes

$$\mathbb{E}_t\tilde{r}_{t+1}^k = \underset{(-)(-)}{a\left(\tilde{\varepsilon}_t, \mathbb{E}_t\tilde{k}_{t+1}\right)} + \tilde{s}_t + b\tilde{k}_t.$$

For every given level of the expected return on capital, the risk-free rate (the return earned by the household on deposits) is $\tilde{s}_t$ per cent lower. Hence, in this *frictional* market, the negative shock generates an additional inward shift of the supply curve as a result of the credit spread rise. In equilibrium, the expected return on capital, $\mathbb{E}_t r_{t+1}^k$, rises further and capital expenditure, $\tilde{k}_t$, falls further than in the frictionless case – and this, at its simplest, is the financial accelerator.

But, why are credit spreads countercyclical? What exactly is the nature of the credit market friction? In this model there are two steps in the intermediation of credit – the process of firms borrowing from banks and the process of banks borrowing from households – either of which could be the source of the friction. The firm might, for example, have an incentive to lie about the return on assets, or the bank might be tempted to abscond with assets, or not be incentivised to exert necessary effort to find good projects.

The next section will formalise these ideas. But, faced with these types of agency problems, the incentives of the borrower (be it the firm or the bank) need to be aligned with the incentives of the creditor (be it the bank or the household). This is achieved when the borrower has 'skin in the game'. In other words, the borrower can no longer rely only on external funds, but must also pledge internal funds. In the frictionless version of this model, the bank was effectively infinitely leveraged

with 100% debt financing. When frictions exist, to make the household willing to supply funds, the bank also needs to provide internal funds.

The key additional state variable in financial friction models is therefore borrowers' net worth (or inside equity). To make profits and accumulate net worth, however, the borrower requires a positive spread between the return on its projects (its assets) and the rate it pays on external finance. When net worth is high, the borrower's incentives are well aligned with that of the household and credit spreads are low. When net worth is low, the benefit from low effort or absconding with funds is relatively high unless credit spreads are high enough such that the opportunity cost of exerting low effort or absconding with assets is also high. As a result, credit spreads are a direct measure of agency costs. Hence the first key additional equilibrium condition in a financial friction model is one that negatively relates current (and future) net worth and current (and future) credit spreads (specific examples of which will be given in the next section).

The second key additional equilibrium condition is the law of motion of net worth, $\tilde{n}_t$. Net worth depends positively on the realised return on capital and positively on net worth in the previous period:

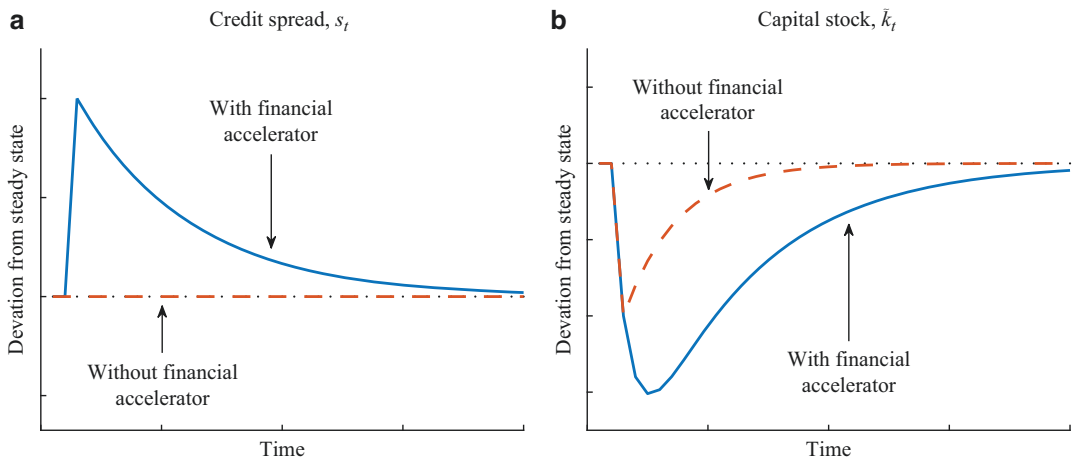
$$\tilde{n}_t = n\left(\tilde{r}_t^k, \tilde{n}_{t-1}\right). \\ (+)\ (+)$$

The bank suffers a hit to net worth whenever the realised return on its assets is below the expected return. This is the case when there is an unexpected negative technology shock, since $r_t^k = \alpha\varepsilon_t k_{t-1}^{\alpha-1}$. The fall in net worth is persistent, propagating the effect of the shock.

In summary, we have established that when microfounded distortions exist in credit markets, credit spreads and borrowers' net worth are negatively related and net worth is procyclical. Thus, we have a model that delivers the countercyclical credit spread needed to generate the financial accelerator.

The financial accelerator sketched above is stylised due to the simplicity of the model. To demonstrate financial accelerator dynamics in a richer DSGE model, Fig. 1 shows the response of the credit spread and capital stock to a negative *capital quality* shock – a shock intended to capture a financial crisis. Specifically, the shock ε_t decreases effective capital from k_t to $\varepsilon_t k_t$ as well as reducing the value of banks' assets.

In Fig. 1a, without the financial accelerator, as in the simple model, there are no agency costs and there is no credit spread. In Fig. 1b, the capital stock falls on impact, but, with the marginal product of capital high as a result, investment rises following the shock and the capital stock recovers quickly. With the financial accelerator, the negative capital quality shock causes an unexpected



Financial Accelerator, Fig. 1 Response to a negative capital quality shock

fall in the return 5 ©Palgrave Macmillan. The New Palgrave Dictionary of Economics. www.dictionaryofeconomics.com. You may not copy or distribute without permission. Licensee: Palgrave Macmillan. on bank assets. Since the bank pays a predetermined risk-free rate on deposits, the bank's net worth gets hit by the shock. Bank net worth falls, exacerbating the friction in the market and driving up the credit spread, as shown in Fig. 1a. This reduces the willingness of households to supply loanable funds, causing investment to fall and the capital stock, as shown in Fig. 1b, to continue falling after the shock (before eventually recovering). Just like the simple model, in this richer DSGE model, the financial accelerator created a large credit spread and an amplified and persistent fall in capital (investment, and output).

Three Microfoundations of the Financial Accelerator

The previous section describes the financial accelerator without reference to a particular microfounded financial friction. This section describes three prominent microfoundations.

Costly State Verification Problem

The costly state verification problem is the microfoundation developed by Bernanke and Gertler (1989). The borrowers facing the friction in this case are risk-neutral entrepreneurs. Entrepreneurs use their own net worth, n_t, and external financing from a bank to purchase capital, k_t at a price q_t for a project. The project is subject to an idiosyncratic productivity shock, $\omega \in (0, \infty)$, the realisation of which is privately observable to the entrepreneur, but only verifiable by the bank by paying a proportional monitoring cost μ. An entrepreneur has an incentive to underreport its gross profit (which is a function of ω). The optimal contract, which ensures truthful reporting by the entrepreneur and minimises the deadweight cost of monitoring, is a standard debt contract. The contract implies a threshold, $\overline{\omega}$. When $\omega \geq \overline{\omega}$, the entrepreneur makes a fixed payment to the bank (and there is no monitoring). When $\omega < \overline{\omega}$, the entrepreneur declares bankruptcy, pays its entire gross profit to the bank, and the bank pays the monitoring cost to audit the entrepreneur. When net worth is low, all else equal, an entrepreneur's incentive to underreport is high. In equilibrium, this causes $\overline{\omega}$, the number of (costly) bankruptcies and the credit spread to all rise. The key equilibrium condition is a trade-off between the credit spread and entrepreneurs' aggregate capital-to-net worth ratio

$$\mathbb{E}_t \tilde{r}^k_{t+1} - \tilde{r}_t = \phi\left(\tilde{q}_t + \tilde{k}_t - \tilde{n}_t\right),$$

where the slope coefficient, ϕ, is a function of the monitoring cost, μ. When $\mu = 0$, then $\phi = 0$ and the model replicates the dynamics of the frictionless economy. Bernanke et al. (1999) showed how variability in the price of capital (through capital adjustment costs) can add additional amplification to the accelerator.

An important technicality of these models is that since the expected discounted return on net worth is above the risk-free rate, it pays for the entrepreneur to always build net worth. With infinitely lived entrepreneurs this would eventually result in the entrepreneurs no longer requiring external finance and the financial accelerator disappearing. To prevent this, there needs to be an exogenous exit rate of entrepreneurs being replaced with new (low net worth) entrepreneurs to ensure that the constraint, in aggregate, continues to bind.

Hold-up Problem

The hold-up problem is the microfoundation developed by Kiyotaki and Moore (1997). Output is produced in two sectors. In the first sector, *productive* agents are impatient and have a constant returns to scale technology. In the second sector, *unproductive* agents are patient and have a decreasing returns to scale technology. The productive agents want to borrow from the unproductive agents but are subject to a friction. Productive agents cannot precommit their human capital, an essential input in production. Thus, they can threaten to repudiate their debt obligations. If they do, the creditors can pay a proportional transaction cost $1 - m$ to repossess the borrower's assets. This generates an endogenous collateral constraint

$b_t \leq m\ \mathbb{E}_t\left(q_{t+1}k_t/r_t\right)$, where b_t is the amount borrowed. In the costly state verification problem, the credit spread was increasing in the relative amount borrowed, since more borrowing required more monitoring. In the hold-up problem the cost of external finance is r_t up to the constraint and then becomes infinite. There are therefore no explicit credit spreads in this model, but the Lagrange multiplier on the collateral constraint can be interpreted as the shadow cost of borrowing. It is the price at which capital can be sold and reallocated – the liquidity of physical capital – that is the key transmission mechanism of shocks. In response to a negative shock, the fire sale of capital from the productive to the unproductive sector depresses asset prices, reducing the collateralisability of assets and hence depressing economic activity. In equilibrium, the productive agents borrow up to the limit and do not consume any of the tradeable output produced. While the productive agents can threaten bankruptcy, in equilibrium this never happens. The problem of productive agents postponing consumption indefinitely also exists in this model, as in Bernanke et al. (1999), and is dealt with by assuming that some output is non-tradeable.

Collateral constraints in the spirit of Kiyotaki and Moore (1997) have been used extensively in the literature with, for example, Iacoviello (2005) using them to study housing dynamics in a new-Keynesian model and Jermann and Quadrini (2012) using collateral constraints and financial shocks to explain the role of debt and equity financing in economic fluctuations.

'Running Away' Moral Hazard Problem

This is the microfoundation developed by Gertler and Karadi (2011). The borrowers facing the friction in this case are financial intermediaries (banks). Households are made up of workers and bankers. Bankers are endowed with an initial net worth from their households and collect deposits from other households to lend to firms. After raising funds, a banker is able to 'run away' with a fraction λ of the bank's total assets. The incentive compatibility constraint is that the fraction of assets with which the banker can run away must be less than the banker's expected discounted terminal net worth. Households therefore only deposit funds at a bank up to the point at which the banker is just indifferent between running away and not. When a banker's current net worth is low, all else equal, its expected discounted terminal net worth is low, and its willingness to run away is high. Thus, in equilibrium, households reduce the quantity of deposits (reducing the absolute value of assets that can be stolen) and credit spreads rise, raising bankers' expected discounted terminal net worth. A contraction in net worth therefore lowers credit creation and raises credit spreads in the economy. The key equilibrium condition is given by

$$\left(\tilde{q}_t + \tilde{k}_t - \tilde{n}_t\right) = \gamma_s\left(\mathbb{E}_t\tilde{r}^k_{t+1} - \tilde{r}_t\right) - \gamma_s\tilde{r}_t + \gamma_\phi\mathbb{E}_t\left(\tilde{q}_{t+1} + \tilde{k}_{t+1} - \tilde{n}_{t+1}\right),$$

where the parameters $\gamma_s, \gamma_\phi > 0$ are functions of λ. Whereas Bernanke et al. (1999) had a static financial friction, with the current credit spread proportional to current leverage, in this setup there is a dynamic financial friction with current leverage increasing in the weighted sum of future credit spreads. As in Kiyotaki and Moore (1997), there is no bankruptcy in equilibrium.

Applications and Challenges

This section discusses the application of financial accelerator models to provide a narrative for the 2007–08 financial crisis and inform monetary and macroprudential policy design, as well as discussing further research challenges.

The Financial Accelerator and the 2007–08 Financial Crisis

The financial crisis was a watershed for the financial accelerator, providing a test case for existing theory and spurring new research. Adrian et al. (2013) assessed the ability of financial friction DSGE models to explain the 2007–08 financial crisis and concluded that models should be able to capture four stylised facts: (1) bank credit falling and credit spreads rising sharply, (2) bond finance increasing, taking up part of the bank credit

supply shortfall, (3) bank equity remaining largely unchanged and (4) bank leverage being highly procyclical.

The simple model described in the earlier section, and most models in the literature, capture stylised fact (1). Few papers, however, capture stylised fact (2), largely because few explicitly model the choice of large firms between bond and bank financing. Adrian et al. (2013) showed that large firms heavily substituted the decline in bank credit with increased bond issuance. This fact helps to identify the collapse in economic activity as a contraction in credit supply by intermediaries rather than a contraction in credit demand by non-financial borrowers. Hence the models of Gertler and Karadi (2011) and Gertler and Kiyotaki (2010), focusing on financial intermediaries, provide a better description of the crisis than earlier models of Bernanke et al. (1999) and Kiyotaki and Moore (1997), focusing on entrepreneurs. However, in stylised models, frictions facing non-financial borrowers can be almost isomorphic to frictions facing intermediaries, and entrepreneurs in many models can be relabelled 'bankers' without much difficulty.

Adrian et al. (2013) argue that standard financial friction models have more difficulty matching stylised facts (3) and (4). To match stylised fact (3) models have often introduced *ad hoc* costs for issuing bank equity. Stylised fact (4), that bank leverage is procyclical, is largely at odds with most financial friction models, as they generate countercyclical leverage. However, Gertler (2013) rejects (3) and (4) as criticisms of current financial accelerator models, arguing that if bank equity and leverage are measured in the data as in the models, then the discrepancy disappears. In models, equity is measured in terms of market values and is highly procyclical, resulting in a countercyclical leverage ratio. In the data, in contrast, equity and assets are measured using a mixture of book value and fair value accounting. And, during liquidity disruptions, even fair value accounting replaces market values with a 'smoothed' value. Thus, bank equity in the data is less procyclical than actual market values would suggest, hence generating procyclical leverage ratios.

A related shortcoming of early generation financial accelerator models was an explanation for why borrowers in 2007 were so leveraged and so reliant on debt. With borrowers assumed only to issue debt in most models, the calibration of a model largely pins down the strength of the financial accelerator. Gertler et al. (2012) extended the model of Gertler and Karadi (2011) by allowing banks to endogenously choose both debt and outside equity financing. Gertler et al. (2012) and de Groot (2014) showed, respectively, how changes in aggregate risk and macroprudential policy, and changes in monetary policy, provide an explanation for the increased reliance of banks on short-term debt financing prior to the crisis and hence an endogenous explanation of why the financial accelerator at that time was so large.

Policy Implications of the Financial Accelerator

The simple financial accelerator model sketched earlier showed technology and capital quality shocks generating inefficient economic fluctuations. An important policy question is whether monetary policy should directly respond to credit market conditions, or respond only in so far as credit market conditions affect output and inflation. Carlstrom et al. (2010), using a hold-up friction, and Fiore and Tristani (2013), using a costly state verification friction, showed, by deriving a utility-based quadratic loss function in a new-Keynesian DSGE model, that welfare is directly affected not just by the usual inflation and output gap volatility terms, but also by a credit spread volatility term. However, the weight on the credit spread term in the welfare approximation is small from a quantitative perspective. Thus, outcomes in response to non-financial shocks would be close to optimal even if monetary policy took no direct account of credit market conditions. In response to technology shocks, near complete inflation stabilisation remains optimal.

With financial shocks, more decisive movements in monetary policy are warranted. However, using monetary policy to offset movements in credit spreads may not be consistent with price stability. This motivates the potential benefits of a second, *macroprudential*, policy instrument with

a financial stability mandate, allowing monetary policy to focus on price stability. Finding the right instrument and coordinating its use with monetary policy is an important research question. Potential instruments include time-varying loan-to-value ratios, liquidity requirements and taxes on borrowing. De Paoli and Paustian (2013) study the coordination problem between monetary and macroprudential policy by deriving a utility-based quadratic loss function in a new-Keynesian DSGE model using a banking friction *à la* Gertler and Karadi (2011). First, they showed that a macroprudential instrument improved outcomes irrespective of potential coordination problems. Second, they showed that while policy set cooperatively and under commitment is optimal, having one instrument act as leader can improve upon policy set non-cooperatively and under discretion (as long as the macroprudential instrument does not affect the economy in too similar a fashion to monetary policy).

Challenges for the Financial Accelerator

The financial accelerator remains an active area of research, and recent contributions have challenged some of the basic assumptions employed in the literature. Dmitriev and Hoddenbagh (2014) and Carlstrom et al. (2016) note that the financial contract between entrepreneurs and banks, specified by Bernanke et al. (1999), was not optimal. First, the original contract assumed that entrepreneurs were myopic, maximising profits today rather than expected discounted terminal net worth. Second, the contract (incorrectly) posited that households want a risk-free return. When the optimal lending contract is derived, with forward-looking entrepreneurs and a state-contingent return for households, the financial accelerator largely disappears. In a similar vein, Candian and Dmitriev (2015) question the commonly used assumption that entrepreneurs are risk-neutral. First, they showed that riskaverse entrepreneurs are more consistent with cross-sectional data. Second, with riskaverse entrepreneurs, they showed that the strength of the financial accelerator was significantly reduced.

Another challenge for financial frictions models is that of identification – the ability to draw inference about the parameters of the model from data. It is usually possible to pin down two friction-relevant parameters by matching steady state moments on leverage and credit spreads. However, insufficient information in time series data causes other parameters to be poorly identified. In estimated versions of Gertler and Karadi (2011), for example, the parameter that governs the life expectancy of bankers is often arbitrarily set at around 10 years. Yet fixing troublesome parameters at arbitrary values can create distortions and lead to false models being selected. With this identification problem it is also difficult to test for time variation in the strength of the financial accelerator.

A third challenge was brought by Chari et al. (2007). Applying a business cycle accounting framework in a canonical business cycle model with wedges, they concluded that the investment wedge – the wedge between the return on capital and the risk-free rate created by financial frictions – did not play a significant role in the Great Depression or postwar recessions, implying that financial accelerator models cannot account for a large share of business cycle dynamics. However, two more recent papers, Jermann and Quadrini (2012) and Christiano et al. (2014), argue that financial frictions combined with financial shocks do play an important role in US business cycles.

Jermann and Quadrini (2012) added a collateral constraint to non-financial borrowers in a quantitative DSGE model and studied the role of financial shocks – shocks to the fraction of assets that can be collateralised for borrowing, m. In line with the suggestion of Chari et al. (2007), Jermann and Quadrini (2012) assumed that firms' labour wage bill also requires financing. With this setup they found that financial shocks play an important role in economic fluctuations, largely because they drive the labour wedge in ways consistent with data. Christiano et al. (2014) estimate a quantitative DSGE model with a costly state verification problem and study the role of *risk* shocks – shocks to the standard deviation, σ, of entrepreneurs' idiosyncratic productivity shocks, $log\ \omega$. They find that risk shocks can account for approximately 60% of US output growth fluctuations.

F

These two papers have shifted the focus from studying the role of the financial accelerator as an amplifier of standard technology and monetary policy shocks to studying the role of shocks originating in the financial sector. The challenge remains to understand whether these new shocks are structural, originating in the financial sector, or are reduced-form representations of important transmission channels lacking in current models.

Conclusion

The theoretical foundations of the financial accelerator mechanism and its qualitative implications are well established. Less agreement – and more scope for future research – exists regarding what are empirically the right financial shocks and frictions and what quantitatively is the role of the financial accelerator in business cycle fluctuations and financial crises. In addition, modelling occasionally binding credit constraints and the full nonlinear implications of financial frictions remains an exciting area of active research.

See Also

- Adjustment Costs
- Calibration
- Credit Cycle
- Great Depression (Mechanisms)
- Great Depression, Monetary and Financial Forces in
- Liquidity Constraints
- Modigliani–Miller Theorem
- Monetary Business Cycle Models (Sticky Prices and Wages)
- Monetary Transmission Mechanism
- Real Business Cycles

Bibliography

Adrian, T., and H.S. Shin. 2014. Procyclical leverage and value-at-risk. *Review of Financial Studies* 27: 373–403.

Adrian, T., P. Colla, and H.S. Shin. 2013. Which financial frictions? Parsing the evidence from the financial crisis of 2007 to 2009. *NBER Macroeconomics Annual* 27(1): 159–214.

Bernanke, B.S., and M. Gertler. 1989. Agency costs, net worth, and business fluctuations. *The American Economic Review* 79(1): 14–31.

Bernanke, B.S., M. Gertler, and S. Gilchrist. 1996. The financial accelerator and the flight to quality. *The Review of Economics and Statistics* 78(1): 1–15.

Bernanke, B.S., M. Gertler, and S. Gilchrist. 1999. The financial accelerator in a quantitative business cycle framework. volume 1, Part C of *Handbook of macroeconomics*, Chapter 21, pp. 1341–1393. Elsevier.

Brock, W., and L. Mirman. 1972. Optimal economic growth and uncertainty: The discounted case. *Journal of Economic Theory* 4(3): 479–513.

Brunnermeier, M.K., T. Eisenbach, and Y. Sannikov. 2013. Macroeconomics with financial frictions: A survey. In *Advances in economics and econometrics: Tenth world congress*, vol. 2. New York: Cambridge University Press.

Candian, G., and M.I. Dmitriev. 2015. Risk aversion and the financial accelerator. Unpublished manuscript.

Carlstrom, C.T., T.S. Fuerst, and M. Paustian. 2010. Optimal monetary policy in a model with agency costs. *Journal of Money, Credit, and Banking* 42: 37–70.

Carlstrom, C.T., T.S. Fuerst, and M. Paustian. 2016. Optimal contracts, aggregate risk, and the financial accelerator. *American Economic Journal: Macroeconomics* 8(1): 119–147.

Chari, V.V., P.J. Kehoe, and E.R. McGrattan. 2007. Business cycle accounting. *Econometrica* 75(3): 781–836.

Christiano, L.J., and D. Ikeda. 2012. *Government policy, credit markets and economic activity.* Unpublished manuscript.

Christiano, L.J., and D. Ikeda. 2013. *Leverage restrictions in a business cycle model.* Unpublished manuscript.

Christiano, L.J., R. Motto, and M. Rostagno. 2014. Risk shocks. *American Economic Review* 104(1): 27–65.

de Groot, O. 2010. *Coordination failure and the financial accelerator.* Unpublished manuscript.

de Groot, O. 2014. The risk channel of monetary policy. *International Journal of Central Banking* 10(2): 115–160.

De Paoli, B., and M. Paustian. 2013. *Coordinating monetary and macroprudential policies.* Unpublished manuscript.

Dmitriev, M., and J. Hoddenbagh. 2014. *The financial accelerator and the optimal lending contract.* Unpublished manuscript.

Fiore, F.D., and O. Tristani. 2013. Optimal monetary policy in a model of the credit channel. *The Economic Journal* 123(571): 906–931.

Fisher, I. 1933. The debt-deflation theory of great depressions. *Econometrica* 1(4): 337–357.

Gertler, M. 2013. Comment on "Which financial frictions? Parsing the evidence from the financial crisis of 2007 to 2009". *NBER Macroeconomics Annual* 27(1): 215–223.

Gertler, M., and P. Karadi. 2011. A model of unconventional monetary policy. *Journal of Monetary Economics* 58(1): 17–34.
Gertler, M., and N. Kiyotaki. 2010. Financial intermediation and credit policy in business cycle analysis. volume 3 of *Handbook of monetary economics*, Chapter 11, pp. 547–599. Elsevier.
Gertler, M., N. Kiyotaki, and A. Queralto. 2012. Financial crises, bank risk exposure and government financial policy. *Journal of Monetary Economics* 59-(Supplement): 17–34.
Goldstein, I., and A. Pauzner. 2005. Demand–deposit contracts and the probability of bank runs. *Journal of Finance* 60(3): 1293–1327.
Hart, O., and J. Moore. 1994. A theory of debt based on the inalienability of human capital. *Quarterly Journal of Economics* 109(4): 841–879.
Iacoviello, M. 2005. House prices, borrowing constraints, and monetary policy in the business cycle. *The American Economic Review* 95(3): 739–764.
Jermann, U., and V. Quadrini. 2012. Macroeconomic effects of financial shocks. *The American Economic Review* 102(1): 238–271.
Keynes, J.M. 1936. *General theory of employment, interest and money.* London: Macmillan.
Kindleberger, C.P. 1978. *Manias, panics and crashes: A history of financial crises*. New York: Basic Books.
Kiyotaki, N., and J. Moore. 1997. Credit cycles. *Journal of Political Economy* 105(2): 211–248.
Minsky, H.P. 1992. *The financial instability hypothesis.* Unpublished manuscript.
Modigliani, F., and M. Miller. 1958. The cost of capital, corporation finance and the theory of investment. *The American Economic Review* 48(3): 261–297.
Quadrini, V. 2011. Financial frictions in macroeconomic fluctuations. *FRB Richmond Economic Quarterly* 97(3): 209–254.
Townsend, R. 1979. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 21(2): 265–293.
Vlcek, J., and S. Roger. 2012. *Macrofinancial modeling at central banks: Recent developments and future directions.* Unpublished manuscript.

Financial Crisis

Charles P. Kindleberger

A financial crisis is defined as a sharp, brief, ultracyclical deterioration of all or most of a group of financial indicators – short-term interest rates, asset (stock, real estate, land) prices, commercial insolvencies and failures of financial institutions (Goldsmith 1982, p. 42). Whereas a boom or bubble is characterized by a rush out of money into real or longer-term financial assets, based on expectations of a continued rise in the price of the asset, financial crisis is characterized by a rush out of the real or long-term financial asset into money, based on the expectation that the price of the asset will decline. Between the boom and a financial crisis may be a period of 'distress' in which the expectation of continued price increases has been eroded, but has not given way to the opposite expectation. Distress may be short or protracted, and may or may not end in crisis.

Bubbles or booms in a single asset, or in widely scattered assets such as Florida or California real estate, supertankers, gold in the 1980s and the like may subside slowly without crisis. The dangers lie in booms or bubbles that have spread from one asset to another, and/or one country to another, and have led to a tautness in the financial structure. In 1825, for example, the boom affected South American government bonds and mining stocks, plus English company securities, led by insurance shares. In 1847, the railway mania was paralleled by a bubble in wheat. The boom following the founding of the German Reich in 1871 affected German and Austrian railways and building, plus lending on United States railway securities. In 1890, the Baring crisis affected Argentina, Brazil, Chile, Australia, South Africa, the United States, France and Italy, which had had booms financed from London and Paris suddenly halted. In the early 1980s, financial distress came from a reversal of the expectation of continued profits in oil and in syndicated bank loans to developing countries. In addition to these loan outlets, financial distress was caused by a boom and crash in farm acreage and in California real estate, plus extensive disintermediation in thrift institutions and financial institutions that had wrongly anticipated a decline in interest rates.

Whether financial distress ends in a financial crisis depends on a variety of factors, including the fragility of the earlier extensions of credit, the speed of the reversal of expectations, the disturbance to confidence produced by some financial

accident (such as a spectacular failure or the revelation of one or more swindles) and the financial community's assurance that in extreme conditions it will be rescued by a lender of last resort.

The function of a domestic lender of last resort was developed in practice in the 18th and early 19th centuries especially by the Bank of England, and rationalized by Walter Bagehot in *Lombard Street* (1873). The task is to halt the rush out of real and long-term financial assets into money by demonstrating to the financial community that there is ample money available. In the crises of 1847, 1857 and 1866 this required suspension of the Bank Act of 1844 limiting the Bank of England's note issue. The lender of last resort ostensibly lent only to solvent institutions on the basis of sound collateral; however, in practice the Bank of England, and especially the Bank of Italy in 1923, 1926 and 1930 made advances on all sorts of assets including many of dubious quality. The assets acquired by the Bank of Italy in its successive interventions to support the weak Italian capital market were consolidated in 1933 into a permanent *Istituto Riconstruzione Italiana* (IRI), patterned somewhat after the 1931 Reconstruction Finance Corporation in the United States.

Other devices historically adopted in financial crises to halt panic liquidation of assets have included: the issue of government securities to merchants against the collateral of their inventories, such securities being sold by the merchant in difficulty; the guaranteeing of the liabilities of distressed commercial or financial institutions as in Hamburg in 1857 and in London in the Baring crisis of 1890; the creation of special intermediaries to add a third signature to bills of exchange to enable them to qualify for ordinary discounting (the French *comptoirs d'escompte* and the *Golddiskontbank* in Germany in 1931); a burst of open-market operations such as undertaken by the Federal Reserve Bank of New York at the end of October and in November 1929 (Kindleberger 1978, ch. 9).

International financial crises in which foreign asset-holders try to dump assets – usually securities or money – to escape from those denominated in a given currency have been less frequently calmed by an international lender of last resort, though this has been done. In the 19th century, the European crises of 1825, 1836, 1839, 1847, 1860, 1890 and 1907 were met by central-bank swaps of gold against silver, or loans of specie or bills of exchange. After World War II a swap network was devised among leading financial centres in which two or more central banks wrote up domestic deposits in favour of the other central bank or banks against a claim in foreign exchange (Coombs 1976). In 1873, 1890 and 1929 international last-resort lending was either absent or inadequate, with the consequence that debt deflation proceeded further and ended in prolonged depression. In 1873 and 1890 there appears to have been no realization of the help that a lender of last resort might have provided. In the 1929 depression, and especially in 1931, Britain was financially too weak to come to the aid of Austria and Germany, and France and the United States failed to recognize the responsibility that had fallen to them (Kindleberger 1973, ch. 14). This view, however, is not universally accepted (Moggridge 1982).

The swap device was not adopted for the debt crises of developing countries in the early 1980s because it was instinctively understood that the resulting claims on them were not certain of ultimate satisfaction. Instead government debts were rescheduled through the so-called Paris Club, and banking claims refunded under the auspices and with the aid of credits from the International Monetary Fund. Since financial crises can occur in a matter of hours, and IMF negotiations are protracted, it has been necessary on occasion, especially for Mexico in 1982, for a 'bridging loan' from a major government or central bank until a more complete settlement could be agreed. These settlements typically required the country being aided to agree to undertake a stringent course of macroeconomic restraint, leading in some instances to internal political unrest (Williamson 1983).

See Also

- ▶ Bank Rate
- ▶ Bubbles
- ▶ Liquidity

Bibliography

Bagehot, W. 1873. *Lombard Street*. Reprinted in *The collected works of Walter Bagehot*, ed. N. St John Stevas. London: The Economist, 1978.
Coombs, C.A. 1976. *The arena of international finance*. New York: Wiley-Interscience.
Goldsmith, R.W. 1982. Comment on Hyman P. Minsky, *The financial instability hypothesis*. In *Financial crises, theory, history and policy*, ed. C.P. Kindleberger and J.-P. Laffargue. Cambridge: Cambridge University Press.
Kindleberger, C.P. 1973. *The world in depression, 1929–1939*. London: Allen Lane.
Kindleberger, C.P. 1978. *Manias, panics and crashes: A history of financial crises*. New York: Basic Books.
Moggridge, D.E. 1982. Policy in the crises of 1920 and 1929. In *Financial crises*, ed. C.P. Kindleberger and J.-P. Laffargue. Cambridge: Cambridge University Press.
Williamson, J. (ed.). 1983. *IMF conditionality*. Washington, DC: Institute for International Economics.

Financial Intermediaries

James Tobin

The tangible wealth of a nation consists of its natural resources, its stocks of goods, and its net claims against the rest of the world. The goods include structures, durable equipment of service to consumers or producers, and inventories of finished goods, raw materials and goods in process. A nation's wealth will help to meet its people's future needs and desires; tangible assets do so in a variety of ways, sometimes by yielding directly consumable goods and services, more often by enhancing the power of human effort and intelligence in producing consumable goods and services. There are many intangible forms of the wealth of a nation, notably the skill, knowledge and character of its population and the framework of law, convention and social interaction that sustains cooperation and community.

Some components of a nation's wealth are appropriable; they can be owned by by governments, or privately by individuals or other legal entities. Some intangible assets are appropriable, notably by patents and copyrights. In a capitalist society most appropriable wealth is privately owned, more than 80 per cent by value in the United States. Private properties are generally transferable from owner to owner. Markets in these properties, *capital markets*, are a prominent feature of capitalist societies. In the absence of slavery, markets in 'human capital' are quite limited.

A person may be wealthy without owning any of the assets counted in appropriable *national wealth*. Instead, a personal wealth inventory would list paper currency and coin, bank deposits, bonds, stocks, mutual funds, cash values of insurance policies and pension rights. These are paper assets evidencing claims of various kinds against other individuals, companies, institutions or governments. In reckoning personal *net worth*, each person would deduct from the value of his total assets the claims of others against him. In 1984 American households' gross holdings of financial assets amounted to about 75 per cent of their net worth, and their net holdings to about 55 per cent (Federal Reserve 1984). If the net worths of all economic units of the nation are added up, paper claims and obligations cancel each other. All that remains, if valuations are consistent and the census is complete, is the value of the national wealth.

If the central government is excluded from this aggregation, *private net worth* – the aggregate net worth of individuals and institutions and subordinate governments (included in the 'private' sector because, lacking monetary powers, they have limited capacities to borrow) – will count not only the national-wealth assets they own but also their net claims against the central government. These include coin and currency, their equivalent in central bank deposit liabilities, and interest-bearing Treasury obligations. If these central government debts exceed the value of its real assets, *private net worth* will exceed national wealth. (However, in reckoning their net worth, private agents may subtract something for the future taxes they expect to pay to service the government's debts. Some economists argue that the subtraction is complete, so that public debt does not count in aggregate private wealth (Barro 1974) while others give reasons the offset is incomplete (Tobin 1980). The issue is not crucial for this essay.)

Outside Assets, Inside Assets and Financial Markets

Private net worth, then, consists of two parts: privately owned items of national wealth, mostly tangible assets, and government obligations. These *outside* assets are owned by private agents not directly but through the intermediation of a complex network of debts and claims, *inside* assets.

Empirical Magnitudes

For the United States at the end of 1984, the value of tangible assets, land and reproducible goods, is estimated at $13.5 trillion, nearly four times the Gross National Product for the year. Of this, $11.2 trillion were privately owned. Adding net claims against the rest of the world and privately owned claims against the federal government gives private net worth of $12.5 trillion, of which only $1.3 trillion represent outside financial assets. The degree of intermediation is indicated by the gross value of financial assets, nearly $14.8 trillion; even if equities in business are regarded as direct titles to real property and excluded from financial assets, the outstanding stock of inside assets is $9.6 trillion. Of these more than half, $5.6 trillion, are claims on financial institutions. The $9.6 million is an underestimate, because many inside financial transactions elude the statisticians. The relative magnitudes of these numbers have changed very little since 1953, when private net worth was $1.27 trillion, gross financial assets $1.35 trillion, $1.05 excluding equities, and GNP was $0.37 trillion (Federal Reserve 1984).

Raymond Goldsmith, who has studied intermediation throughout a long and distinguished career and knows far more about it than anyone else, has estimated measures of intermediation for many countries over long periods of time (1969, 1985). Here is his own summary:

> The creation of a modern financial superstructure, not in its details but in its essentials, was generally accomplished at a fairly early stage of a country's economic development, usually within five to seven decades from the start of modern economic growth. Thus it was essentially completed in most now-developed countries by the end of the 19th century or the eve of World War I, though somewhat earlier in Great Britain. During this period the financial interrelations ratio, the quotient of financial and tangible assets, increased fairly continuously and sharply. Since World War I or the Great Depression, however, the ratio in most of these countries has shown no upward trend, though considerable movements have occurred over shorter periods, such as sharp reductions during inflations; and though significant changes have taken place in the relative importance of the various types of financial institutions and of financial instruments.
>
> Among less developed countries, on the other hand, the financial interrelations ratio has increased substantially, particularly in the postwar period, though it generally is still well below the level reached by the now-developed countries early in the 20th century.

Goldsmith finds that a ratio of the order of unity is characteristic of financial maturity, as is illustrated by the figures for the United States given above (1985, pp. 2–3).

Goldsmith finds also that the relative importance of financial institutions, especially non–banks, has trended upwards in most market economies but appears to taper off in mature systems. Institutions typically hold from a quarter to a half of all financial instruments. Ratios around 0.40 were typical in 1978, but there is considerably more variation among countries than in the financial interrelations ratio. The United States, at 0.27, is on the low side, probably because of its many well-organized financial markets (1985, Table 47, p. 136).

The volume of gross financial transactions is mind-boggling. The GNP velocity of the money stock in the United States is 6 or 7 per year; if intermediate as well as final transactions for goods and services are considered, the turnover may be 20 or 30 per year. But demand deposits turn over 500 times a year, 2500 times in New York City banks, indicating that most transactions are financial in nature. The value of stock market transactions alone in the United States is one third of the Gross National Product; an average share of stock changes hands every nineteen months. Gross foreign exchange transactions in United States dollars are estimated to be hundreds of billions of dollars every day. 'Value added' in the financial services industries amounts to 9 per cent of United States GNP (Tobin 1984).

Outside and Inside Money

The outside/inside distinction is most frequently applied to money. *Outside money* is the monetary debt of the government and its central bank, currency and central bank deposits, sometimes referred to as 'base' or 'high-powered' money. *Inside money*, 'low-powered', consists of private deposit obligations of other banks and depository institutions in excess of their holdings of outside money assets. Just which kinds of deposit obligations count as 'money' depends on definitions, of which there are several, all somewhat arbitrary. Outside money in the United States amounted to $186 billion at the end of 1983, of which $36 billion was held as reserves by banks and other depository institutions; the remaining $150 billion was held by other private agents as currency. The total money stock M1, currency in public circulation plus checkable deposits, was $480 billion. Thus inside M1 was $294 billion, more than 60 per cent of the total.

Financial Markets, Organized and Informal

Inside assets and debts wash out in aggregative accounting; one person's asset is another's debt. But for the functioning of the economy, the inside network is of great importance. *Financial markets* allow inside assets and debts to be originated and to be exchanged at will for each other and for outside financial assets. These markets deal in paper contracts and claims. They complement the markets for real properties. Private agents often borrow to buy real property and pledge the property as security; households mortgage new homes, businesses incur debt to acquire stocks of materials or goods-in-process or to purchase structures and equipment. The term *capital markets* covers both financial and property markets. *Money markets* are financial markets in which short-term debts are exchanged for outside money.

Many of the assets traded in financial markets are promises to pay currency in specified amounts at specified future dates, sometimes conditional on future events and circumstances. The currency is not always the local currency; obligations denominated in various national currencies are traded all over the world. Many traded assets are not denominated in any future monetary unit of account: equity shares in corporations, contracts for deliveries of commodities – gold, oil, soy beans, hog bellies. There are various hybrid assets: preferred stock gives holders priority in distributions of company profits up to specified pecuniary limits; convertible debentures combine promises to pay currency with rights to exchange the securities for shares.

Capital markets, including financial markets, take a variety of forms. Some are highly organized auction markets, the leading real-world approximations to the abstract perfect markets of economic theory, where all transactions occurring at any moment in a commodity or security are made at a single price and every agent who wants to buy or sell at that price is accommodated. Such markets exist in shares, bonds, overnight loans of outside money, standard commodities, and foreign currency deposits, and in futures contracts and options for most of the same items.

However, many financial and property transactions occur otherwise, in direct negotiations between the parties. Organized open markets require large tradable supplies of precisely defined homogeneous commodities or instruments. Many financial obligations are one of a kind, the promissory note of a local business proprietor, the mortgage on a specific farm or residence. The terms, conditions, and collateral are specific to the case. The habit of referring to classes of heterogeneous negotiated transactions as 'markets' is metaphorical, like the use of the term 'labour market' to refer to the decentralized processes by which wages are set and jobs are filled, or 'computer market' to describe the pricing and selling of a host of differentiated products. In these cases the economists' faith is that the outcomes are 'as if' the transaction occurred in perfect organized auction markets.

Financial Enterprises and Their Markets

Financial intermediaries are enterprises in the business of buying and selling financial assets. The accounting balance sheet of a financial intermediary is virtually 100 per cent paper on both

sides. The typical financial intermediary owns relatively little real property, just the structures, equipment, and materials necessary to its business. The equity of the owners, or the equivalent capital reserve account for mutual, cooperative, nonprofit, or public institutions, is small compared to the enterprises' financial obligations.

Financial intermediaries are major participants in organized financial markets. They take large asset positions in market instruments; their equities and some of their liabilities, certificates of deposit or debt securities, are traded in those markets. They are not just middlemen like dealers and brokers whose main business is to execute transactions for clients.

Fnancial intermediaries are the principal makers of the informal financial markets discussed above. Banks and savings institutions hold mortgages, commercial loans, and consumer credit; their liabilities are mainly checking accounts, savings deposits, and certificates of deposit. Insurance companies and pension funds negotiate private placements of corporate bonds and commercial mortgages; their liabilities are contracts with policy-holders and obligations to future retirees. Thus financial intermediaries do much more than participate in organized markets. If financial intermediaries confined themselves to repackaging open market securities for the convenience of their creditors, they would be much less significant actors on the economic scene.

Financial businesses seek customers, both lenders and borrowers, not only by interest rate competition but by differentiating and advertising their 'products'. Financial products are easy to differentiate, by variations in maturities, fees, auxiliary services, office locations and hours of business, and many other features. As might be expected, non-price competition is especially active when prices, in this case interest rates, are fixed by regulation or by tacit or explicit collusion. But the industry is by the heterogeneous nature of its products monopolistically competitive; non-price competition flourishes even when interest rates are free to move. The industry shows symptoms of 'wastes of monopolistic competition'. Retail offices of banks and savings institutions cluster like competing gasoline stations. Much claimed product differentiation is trivial and atmospheric, emphasized and exaggerated in advertising.

Financial intermediaries cultivate long-term relationships with customers. Even in the highly decentralized financial system of the United States, local financial intermediaries have some monopoly power, some clienteles who will stay with them even if their interest rates are somewhat less favourable than those elsewhere. Since much business is bilaterally negotiated, there are ample opportunities for price discrimination. The typical business customer of a bank is both a borrower and a depositor, often simultaneously. The customer 'earns' the right for credit accommodation when he needs it by lending surplus funds to the same bank when he has them. The same reciprocity occurs between credit unions and mutual savings institutions and some of their members. Close ties frequently develop between a financial intermediary and non-financial businesses whose sales depend on availability of credit to their customers, for example between automobile dealers and banks. Likewise, builders and realtors have funded and controlled many savings and loan associations in order to facilitate mortgage lending to home buyers.

Financial intermediaries balance the credit demands they face with their available funds by adjusting not only interest rates but also the other terms of loans. They also engage in quantitative rationing, the degree of stringency varying with the availability and costs of funds to the intermediary. Rationing occurs naturally as a by-product of lending decisions made and negotiated case by case. Most such loans require collateral, and the amount and quality of the collateral can be adjusted both to individual circumstances and to overall market conditions. Borrowers are classified as to riskiness and charged rates that vary with their classification.

United States commercial banks follow the 'prime rate convention'. One or another of the large banks acts as price leader and sets a rate on six-month commercial loans for its prime quality borrowers. If other large banks agree, as is usually the case, they follow, and the rate becomes standard for the whole industry until one of the

leading banks decides another change is needed to stay in line with open-market interest rates. Loan customers are rated by the number of half-points above prime at which they will be accommodated. Of course, some applications for credit are just turned away. One mechanism of short-term adjustment to credit market conditions is to stiffen or relax the risk classifications of customers, likewise to deny credit to more or fewer applicants. Similar mechanisms for rationing help to equate demands to supplies of home mortgage finance and consumer credit.

The Functions of Financial Markets and Intermediary Institutions

Intermediation, as defined and described above, converts the outside privately owned wealth of the economy into the quite different forms in which its ultimate owners hold their accumulated savings. Financial markets alone accomplish considerable intermediation, just by facilitating the origination and exchange of inside assets. Financial intermediaries greatly extend the process, adding 'markets' that would not exist without them, and participating along with other agents in other markets, organized or informal.

What economic functions does intermediation in general perform? What do inside markets add to markets in the basic outside assets? What functions does institutional intermediation by financial intermediaries perform beyond those of open markets in financial instruments? Economists characteristically impose on themselves questions like these, which do not seem problematic to lay practitioners. Economists start from the presumption that financial activities are epiphenomena, that they create a veil obscuring to superficial observers an underlying reality which they do not affect. The celebrated Modigliani–Miller theorem (1958), generalized beyond the original intent of the authors, says so. With its help the sophisticated economist can pierce the veil and see that the values of financial assets are just those of the outside assets to which they are ultimately claims, no matter how circuitous the path from the one to the other.

However, economists also understand how the availability of certain markets alters, usually for the better, the outcomes prevailing in their absence. For a primitive illustration, consider the functions of inside loan markets as brilliantly described by Irving Fisher (1930). Each household has an inter-temporal utility function in consumptions today and at future times, a sequence of what we now would call dated 'endowments' of consumption, and an individual 'backyard' production function by which consumption less than endowment at any one date can be transformed into consumption above endowment at another date. Absent the possibility of intertemporal trades with others, each household has to do its best on its own; its best will be to equate its marginal rate of substitution in utility between any two dates with its marginal rate of transformation in production between the same dates, with the usual amendments for corner solutions. The gains from trade, i.e., in this case from auction markets in inter-household lending and borrowing, arise from differences among households in those autarkic rates of substitution and transformation. They are qualitatively the same as those from free contemporaneous trade in commodities between agents or nations.

The introduction of consumer loans in this Fisherian model will alter the individual and aggregate paths of consumption and saving. It is not possible to say whether it will raise or lower the aggregate amount of capital, here in the sense of labour endowments in the process of producing future rather than current consumable output. In either case it is likely to be a Pareto-optimal improvement, although even this is not guaranteed *a priori*.

Similar argument suggests several reasons why ultimately savers, lenders and creditors prefer the liabilities of financial intermediaries not only to direct ownership of real property but also to the direct debt and equity issues of investors, borrowers and debtors:

Convenience of Denomination

Issuers of securities find it costly to cut their issues into the variety of small and large denominations savers find convenient and commensurate to their

means. The financial intermediary can break up large-denomination bonds and loans into amounts convenient to small savers, or combine debtors' obligations into large amounts convenient to the wealthy. Economies of scale and specialization in financial transactions enable financial intermediaries to tailor assets and liabilities to the needs and preferences of both lenders and borrowers. This service is especially valuable for agents on both sides whose needs vary in amount continuously; they like deposit accounts and credit lines whose use they can vary at will on their own initiative.

Risk Pooling, Reduction and Allocation

The risks incident to economic activities take many forms. Some are nation-wide or world-wide – wars and revolutions, shifts in international comparative advantage, government fiscal and monetary policies, prices and supplies of oil and other basic materials. Some are specific to particular enterprises and technologies – the capacity and integrity of managers, the qualities of new products, the local weather. A financial intermediary can specialize in the appraisal of risks, especially specific risks, with expertise in the gathering and interpretation of information costly or unavailable to individual savers. By pooling the funds of its creditors, the financial intermediary can diversify away risks to an extent that the individual creditors cannot, because of the costs of transactions as well as the inconvenience of fixed lumpy denominations.

According to Joseph Schumpeter ([1911] 1934, pp. 72–4), bankers are the gatekeepers – Schumpeter's word is 'ephor' – of capitalist economic development; their strategic function is to screen potential innovators and advance the necessary purchasing power to the most promising. They are the source of purchasing power for investment and innovation, beyond the savings accumulated from past economic development. In practice, the cachet of a banker often enables his customer also to obtain credit from other sources or to float paper in open markets.

Maturity Shifting

A financial intermediary typically reconciles differences among borrowers and lenders in the timing of payments. Bank depositors want to commit funds for shorter times than borrowers want to have them. Business borrowers need credit to bridge the time gap between the inputs to profitable production and their output and sales. This source of bank business is formally modelled by Diamond and Dybvig (1983). The bank's scale of operations enables it to stagger the due dates of, say, half-year loans so as to accommodate depositors who want their money back in three months or one month or on demand. The reverse maturity shift may occur in other financial intermediaries. An insurance company or pension fund might invest short-term the savings its policy-owners or future pensioners will not claim for many years.

Transforming Illiquid Assets into Liquid Liabilities

Liquidity is a matter of degree. A perfectly liquid asset may be defined as one whose full present value can be realized, i.e., turned into purchasing power over goods and services, immediately. Dollar bills are perfectly liquid, and so for practical purposes are demand deposits and other deposits transferable to third parties by check or wire. Liquidity in this sense does not necessarily mean predictability of value. Securities traded on well organized markets are liquid. Any person selling at a given time will get the same price whether he decided and prepared to sell a month before or on the spur of the moment. But the price itself can vary unpredictably from minute to minute. Contrast a house, neither fully liquid nor predictable in value. Its selling proceeds at this moment are likely to be greater the longer it has been on the market. Consider the six-month promissory note of a small business proprietor known only to his local banker. However sure the payment on the scheduled date, the note may not be marketable at all. If the lender wants to realize its value before maturity, he will have to find a buyer and negotiate. A financial intermediary holds illiquid assets while its liabilities are liquid, and holds assets unpredictable in value while it guarantees the value of its liabilities. This is the traditional business of commercial banks, and the reason for the strong and durable relations of banks and their customers.

Substitution of Inside for Outside Assets

What determines the aggregate liabilities and assets of financial intermediaries? What determines the gross aggregate of inside assets generated by financial markets in general, including open markets as well as financial intermediaries? How can the empirical regularities found by Goldsmith, cited above, be explained?

Economic theory offers no answers to these questions. The differences among agents that invite mutually beneficial transactions, like those discussed above, offer opportunities for inside markets. Theory can tell us little *a priori* about the size of such differences. Moreover, markets are costly to operate, whether they are organized auction markets in homogeneous instruments or the imperfect 'markets' in heterogeneous contracts in which financial intermediaries are major participants. Society cannot afford all the markets that might exist in the absence of transactions costs and other frictions, and theory has little to say on which will arise and survive.

The macroeconomic consequence of inside markets and financial intermediaries is generally to provide substitutes for outside assets and thus to economize their supplies. That is, the same microeconomic outcomes are achievable with smaller supplies of one or more of the outside assets than in the absence of intermediation. The way in which intermediation mobilizes the surpluses of some agents to finance the deficits of others is the theme of the classic influential work of Gurley and Shaw (1960).

Consider, for example, how commercial banking diminishes the need of business firms for net worth invested in inventories, by channelling the seasonal cash surpluses of some firms to the contemporaneous seasonal deficits of others. Imagine two firms A and B with opposite and complementary seasonal zigzag patterns. A needs \$2 in cash at time zero to buy inputs for production in period 1 sold for \$2; the pattern repeats in 3, 4,... B needs \$2 in cash at time 1 to buy inputs for production in period 2 sold for \$2 in period 3, and so on in 4, 5,.... In the absence of their commercial bank, A and B each need \$2 of net worth to carry on business; from period to period each alternates holding it in cash and in goods-in-process. between them the two firms always are holding \$2 of currency and \$2 of inventories. B enters the bank and lends A half the \$2 he needs to carry his inventory in period 1; A repays the loan from sales proceeds the next period, 2; the bank now lends \$1 to B, ... A and B now need only \$1 of currency; each has on average net worth of \$1.50 – \$2 and \$1 alternating; as before they are together always holding \$2 of inventories. Moreover, with a steady deposit of \$2 from a third party, the bank could finance both businesses completely; they would need no net worth of their own. The example is trivial, but commercial banking proper can be understood as circulation of deposits and loans among businesses and as a revolving fund assembled from other sources and lent to businesses.

As a second primitive example, consider the effects of introducing markets that enable risks to be borne by those households more prepared to take them. Suppose that of two primary outside assets, currency and tangible capital, the return on the latter has the greater variance. Individuals who are risk neutral will hold all their wealth (possibly excepting minimal transactions balances of currency) in capital as long as its expected return exceeds the expected real return on currency. If these more adventurous households are not numerous and wealthy enough to absorb all the capital, the expected return on capital will have to exceed that on currency enough to induce risk-averse wealth-owners to hold the remainder. In this equilibrium the money price of capital and its mean real return are determined so as to allocate the two assets between the two kinds of households. Now suppose that the risk-neutral households can borrow from the risk-averse types, most realistically via financial intermediaries, and that the latter households regard those debts as close substitutes for currency, indeed as inside money if intermediation by financial intermediaries is involved. The inside assets do double duty, providing the services and security of money to those who value them while enabling the more adventurous to hold capital in excess of their own net worth. As a result, the private sector as a whole will want to hold a larger proportion of its wealth in capital at any given expected real return on

capital. In equilibrium, the aggregate capital stock will be larger and its expected return, equal to its marginal productivity in a steady state, will be lower than in the absence of intermediation.

Intermediation can diminish the private sector's need not just for outside money but for net worth and tangible capital. These economies generally require financial markets in which financial intermediaries are major participants, because they involve heterogeneous credit instruments and risk pooling. In the absence of home mortgages, consumer credit, and personal loans for education, young households would not be able to spend their future wages and salaries until they receive them. Constraints on borrowing against future earnings make the age-weighted average net non-human wealth of the population greater, but the relaxation of such liquidity constraints increases household welfare. Financial intermediaries invest the savings of older and more affluent households in loans to their younger and less wealthy contemporaries; otherwise those savings would go into outside assets. Likewise insurance makes it unnecessary to accumulate savings as precaution against certain risks, for example the living and medical expenses of unusual longevity. It is an all too common fallacy to assume that arrangements that increase aggregate savings and tangible wealth always augment social welfare.

Deposit Creation and Reserve Requirements

The substitution of inside money for outside money is the familiar story of deposit creation, in which the banking system turns a dollar of base or 'high-powered' money into several dollars of deposits. The extra dollars are inside or 'low-powered' money. The banks need to hold only a fraction k, set by law or convention or prudence, of their deposit liabilities as reserves in base money. In an equilibrium in which they hold no excess reserves their deposits will be a multiple $1/k$ of their reserves; they will have created $(1 - k)/k$ dollars of substitute money.

A key step in this process is that any bank with excess reserves makes a roughly equal amount of additional loans, crediting the borrowers with deposits. As the borrowers draw checks, these new deposits are transferred to other accounts, most likely in other banks. As deposits move to other banks, so do reserves, dollar for dollar. But now those banks have excess reserves and act in like manner. The process continues until all banks are 'loaned up', i.e. deposits have increased enough so that the initial excess reserves have become reserves that the banks require or desire.

The textbook fable of deposit creation does not do justice to the full macroeconomics of the process. The story is incomplete without explaining how the public is induced to borrow more and to hold more deposits. The borrowers and the depositors are not the same public. No one borrows at interest in order to hold idle deposits. To attract additional borrowers, banks must lower interest rates or relax their collateral requirements or their risk standards. The new borrowers are likely to be businesses that need bank credit to build up inventories of materials or goods in process. The loans lead quickly to additional production and economic activity. Or banks buy securities in the open market, raising their prices and lowering market interest rates. The lower market rates may encourage businesses to float issues of commercial paper, bonds or stocks, but the effects of investment in inventories or plant and equipment are less immediate and less potent than the extension of bank credit to a business otherwise held back by illiquidity. In either case, lower interest rates induce other members of the public, those who indirectly receive the loan disbursements or those who sell securities to banks, to hold additional deposits. They will be acquiring other assets as well, some in banks, some in other financial intermediaries, some in open financial markets. Lower interest rates may also induce banks themselves to hold extra excess reserves.

Interest rates are not the only variables of adjustment. Nominal incomes are rising at the same time, in some mixture of real quantities and prices depending on macroeconomic circumstances. The rise in incomes and economic activities creates new needs for transactions balances of money. Thus the process by which excess reserves are absorbed entails changes in interest

rates, real economic activity, and prices in some combination. It is possible to describe scenarios in which the entire ultimate adjustment is in one of these variables. Wicksell's cumulative credit expansion, which in the end just raises prices, is a classic example.

Do banks have a unique magic by which asset purchases generate their own financing? Is the magic due to the 'moneyness' of the banks' liabilities? The preceding account indicates it is not magic but reserve requirements. Moreover, a qualitatively similar story could be told if reserve requirements were related to bank assets or non-monetary liabilities and even if banks happened to have no monetary liabilities at all. In the absence of reserve requirements aggregate bank assets and liabilities, relative to the size of the economy, would be naturally limited by public supplies and demands at interest rates that cover banks' costs and normal profits. If, instead of banks, savings institutions specializing in mortgage lending were subject to reserve requirements, their incentives to minimize excess reserves would inspire a story telling how additional mortgage lending brings home savings deposits to match (Tobin 1963).

Risks, Runs and Regulations

Some financial intermediaries confine themselves to activities that entail virtually no risk either to the institution itself or to its clients. An open-end mutual fund or unit trust holds only fully liquid assets traded continuously in organized markets. It promises the owners of its shares payment on demand at their pro rata net value calculated at the market prices of the underlying assets – no more, no less. The fund can always meet such demands by selling assets it holds. The shareowners pay in one way or another an agreed fee from the services of the fund – the convenience and flexibility of denomination, the bookkeeping, the transactions costs, the diversification, the expertise in choosing assets. The shareowners bear the market risks on the fund's portfolio – no less and, assuming the fund is honest, no more. Government regulations are largely confined to those governing all public security issues, designed to protect buyers from deceptions and insider manipulations. In the United States regulation of this kind is the province of the federal Securities and Exchange Commission.

Most financial intermediaries do take risks. The risks are intrinsic to the functions they serve and to the profit opportunities attracting financial entrepreneurs and investors in their enterprises. For banks and similar financial intermediaries, the principal risk is that depositors may at any time demand payments the institution can meet, if at all, only at extraordinary cost. Many of the assets are illiquid, unmarketable. Others can be liquidated at short notice only at substantial loss. In some cases, bad luck or imprudent management brings insolvency; the institution could never meet its obligations no matter how long its depositors and other creditors wait. In other cases, the problem is just illiquidity; the assets would suffice if they could be held until maturity, until buyers or lenders could be found, or until normal market conditions returned.

Banks and other financial intermediaries hold reserves, in currency or its equivalent, deposits in central banks, or in other liquid forms as precaution against withdrawals by their depositors. For a single bank, the withdrawal is usually a shift of deposits to other banks or financial intermediaries, arising from a negative balance in interbank clearings of checks or other transfers to third parties at the initiative of depositors. For the banking system, as a whole, withdrawal is a shift by the public from deposits to currency.

'Withdrawals' may in practice include the exercise of previously agreed borrowing rights. Automatic overdraft privileges are more common in other countries, notably the United Kingdom and British Commonwealth nations, than in the United States. They are becoming more frequent in the United States as an adjunct of bank credit cards. Banks' business loan customers often have explicit or implicit credit lines on which they can draw on demand.

Unless financial intermediaries hold safe liquid assets of predictable value matched in maturities to their liabilities – in particular, currency or equivalent against all their demand

obligations – they and their creditors can never be completely protected from withdrawals. The same is true of the banking system as a whole, and of all intermediaries other than simple mutual funds. 'Runs', sudden, massive, and contagious withdrawals, are always possible. They destroy prudent and imprudent institutions alike, along with their depositors and creditors. Of course, careful depositors inform themselves about the intermediaries to which they entrust their funds, about their asset portfolios, policies and skills. Their choices among competing depositories provide some discipline, but it can never be enough to rule out disasters. What the most careful depositor cannot foresee is the behaviour of other depositors, and it is rational for the well-informed depositor of a sound bank to withdraw funds if he believes that others are doing so or are about to do so.

Governments generally regulate the activities of banks and other financial intermediaries in greater detail than they do nonfinancial enterprises. The basic motivations for regulation appear to be the following:

It is costly, perhaps impossible, for individual depositors to appraise the soundness and liquidity of financial institutions and to estimate the probabilities of failures even if they could assume that other depositors would do likewise. It is impossible for them to estimate the probabilities of 'runs'. Without regulation, the liabilities of suspect institutions would be valued below par in check collections. Prior to 1866 banks in the United States were allowed to issue notes payable to bearers on demand, surrogates for government currency. The notes circulated at discounts varying with the current reputations of the issuers. A system in which transactions media other than government currency continuously vary in value depending on the issuer is clumsy and costly.

The government has obligation to provide at low social cost an efficient system of transactions media, and also a menu of secure and convenient assets for citizens who wish to save in the national monetary unit of account. Those transactions media and saving assets can be offered by banks and other financial intermediaries, in a way that retains most of the efficiencies of decentralization and competition, if and only if government imposes some regulations and assumes some residual responsibilities. The government's role takes several forms.

Reserve Requirements

An early and obvious intervention was to require banks to hold reserves in designated safe and liquid forms against their obligations, especially their demand liabilities. Left to themselves, without such requirements, some banks might sacrifice prudence for short-term profit. Paradoxically, however, required reserves are not available for meeting withdrawals unless the required ratio is 100 per cent. If the reserve requirement is 10 per cent of deposits, then withdrawal of one dollar from a bank reduces its reserve holdings by one dollar but its reserve requirement by only ten cents. Only excess reserves or other liquid assets are precautions against withdrawals. The legal reserve requirement just shifts the bank's prudential calculation to the size of these secondary reserves. Reserve requirements serve functions quite different from their original motivation. In the systems that use them, notably the United States, they are the fulcrum for central bank control of economy-wide monetary conditions. (They are also an interest-free source of finance of government debt, but in the United States today this amounts to only $45 billion of a total debt to the public of $1700 billion.)

Last-resort Lending

Banks and other financial intermediaries facing temporary shortages of reserves and secondary reserves of liquid assets can borrow them from other institutions. In the United States, for example, the well-organized market for 'federal funds' allows banks short of reserves to borrow them overnight from other banks. Or banks can gain reserves by attracting more deposits, offering higher interest rates on them than depositors are getting elsewhere. These ways of correcting reserve positions are not available to troubled banks, suspected of deep-rooted problems of liquidity or solvency or both, for example bad loans. Nor will they meet a system-wide run from liabilities of banks and other financial intermediaries into currency.

Banks in need of reserves can also borrow from the central bank, and much of this borrowing is routine, temporary, and seasonal. Massive central bank credit is the last resort of troubled banks which cannot otherwise satisfy the demands of their depositors without forced liquidations of their assets. The government is the ultimate supplier of currency and reserves in aggregate. The primary *raison d'être* of the central bank is to protect the economy from runs into currency. System-wide shortages of currency and reserves can be relieved not only by central bank lending to individual banks but by central bank purchases of securities in the open market. The Federal Reserve's inability or unwillingness – which it was is still debated – to supply the currency bank depositors wanted in the early 1930s led to disastrous panic and epidemic bank failures. No legal or doctrinal obstacles would now stand in the way of such a rescue.

Deposit Insurance

Federal insurance of bank deposits in the United States has effectively prevented contagious runs and epidemic failures since its enactment in 1935. Similar insurance applies to deposits in savings institutions. In effect, the federal government assumes a contingent residual liability to pay the insured deposits in full, even if the assets of the financial intermediary are permanently inadequate to do so. The insured institutions are charged premiums for the service, but the fund in which they are accumulated is not and cannot be large enough to eliminate possible calls on the Treasury. Although the guarantees are legally limited to a certain amount, now $100,000, per account, in practice depositors have eventually recovered their full deposits in most cases. Indeed the guarantee seems now to have been extended *de facto* to all deposits, at least in major banks.

Deposit insurance impairs such discipline as surveillance by large depositors might impose on financial intermediaries; instead the task of surveillance falls on the governmental insurance agencies themselves (in the United States the Federal Deposit Insurance Corporation and the Federal Savings and Loan Insurance Corporation) and on other regulatory authorities (the United States Comptroller of the Currency, the Federal Reserve, and various state agencies). Insurance transfers some risks from financial intermediary depositors and owners to taxpayers at large, while virtually eliminating risks of runs. Those are risks we generate ourselves; they magnify the unavoidable natural risks of economic life. Insurance is a mutual compact to enable us to refrain from *sauve qui peut* behaviour that can inflict grave damage on us all. Formally, an uninsured system has two equilibria, a good one with mutual confidence and a bad one with runs. Deposit insurance eliminates the bad one (Diamond and Dybvig 1983).

One hundred per cent reserve deposits would, of course, be perfectly safe – that is, as safe as the national currency – and would not have to be insured. Those deposits would in effect *be* currency, but in a secure and conveniently checkable form. One can imagine a system in which banks and other financial intermediaries offered such accounts, with the reserves behind them segregated from those related to the other business of the institution. That other business would include receiving deposits which required fractional or zero reserves and were insured only partially, if at all. The costs of the 100 per cent reserve deposit accounts would be met by service charges, or by government interest payments on the reserves, justified by the social benefits of a safe and efficient transactions medium. The burden of risk and supervision now placed on the insuring and regulating agencies would be greatly relieved. It is, after all, historical accident that supplies of transactions media in modern economies came to be byproducts of banking business and vulnerable to its risks.

Government may insure financial intermediaries loans as well as deposits. Insurance of home mortgages in the United States not only has protected the institutions that hold them and their depositors but has converted the insured mortgages into marketable instruments.

Balance Sheet Supervision

Government surveillance of financial intermediaries limits their freedom of choice of assets and liabilities, in order to limit the risks to depositors

and insurers. Standards of adequacy of capital – owners' equity at risk in the case of private corporations, net worth in the case of mutual and other nonprofit forms of organization – are enforced for the same reasons. Periodic examinations check the condition of the institution, the quality of its loans, and the accuracy of its accounting statements. The regulators may close an institution if further operation is judged to be damaging to the interests of the depositors and the insurer.

Legislation which regulates financial intermediaries has differentiated them by purpose and function. Commercial banks, savings institutions, home building societies, credit unions, and insurance companies are legally organized for different purposes. They are subject to different rules governing the nature of their assets. For example, home building societies – savings and loan associations in the United States – have been required to keep most of their asset portfolios in residential mortgages. Restrictions of this kind mean that when wealth-owners shift funds from one type of financial intermediary to another, they alter relative demands for assets of different kinds. Shifts of deposits from commercial banks to building societies would increase mortgage lending relative to commercial lending. Regulations have also restricted the kinds of liabilities allowed various types of financial intermediary. Until recently in the United States, only banks were permitted to have liabilities payable on demand to third parties by check or wire. Currently deregulation is relaxing specialized restrictions on financial intermediary assets and liabilities and blurring historical distinctions of purpose and function.

Interest Ceilings

Government regulations in many countries set ceilings on the interest rates that can be charged on loans and on the rates that can be paid on deposits, both at banks and at other financial intermediaries. In the United States the Banking Act of 1935 prohibited payment of interest on demand deposits. After the second world war effective ceilings on savings and time deposits in banks and savings institutions were administratively set, and on occasion changed, by federal agencies. Under legislation of 1980, these regulations are being phased out.

> The operating characteristics of a system of financial intermediaries in which interest rates on deposits of various types, as well as on loans, are set by free competition are quite different from those of a system in which financial intermediary rates are subject to legal ceilings or central bank guidance, or set by agreement among a small number of institutions. For example, when rates on deposits are administratively set, funds flow out of financial intermediaries when open market rates rise and return to financial intermediaries when they fall. These processes of 'disintermediation' and 're-intermediation' are diminished when financial intermediary rates are free to move parallel to open market rates. Likewise flows between different financial intermediaries due to administratively set rate differences among them are reduced when they are all free to compete for funds.

A regime with market-determined interest rates on moneys and near-moneys has significantly different macroeconomic characteristics from a regime constrained by ceilings on deposit interest rates. Since the opportunity cost of holding deposits is largely independent of the general level of interest rates, the 'LM' curve is steeper in the unregulated regime. Both central bank operations and exogenous monetary shocks could be expected to have larger effects on nominal income, while fiscal measures and other shocks to aggregate demand for goods and services would have smaller effects (Tobin 1983).

Entry, Branching, Merging

Entry into regulated financial businesses is generally controlled, as are establishing branches or subsidiaries and merging of existing institutions. In the United States, charters are issued either by the federal government or by state governments, and regulatory powers are also divided. Until recently banks and savings institutions, no matter by whom chartered, were not allowed to operate in more than one state. This rule, combined with various restrictions on branches within states, gave the United States a much larger number of distinct financial enterprises, many of them very small and very local, than is typical in other countries. The prohibition of interstate operations is

now being eroded and may be effectively eliminated in the next few years.

Deregulation has been forced by innovations in financial technology that made old regulations either easy hurdles to circumvent or obsolete barriers to efficiency. New opportunities not only are breaking down the walls separating financial intermediaries of different types and specializations. They are also bringing other businesses, both financial and nonfinancial, into activities previously reserved to regulated financial institutions. Mutual funds and brokers offer accounts from which funds can be withdrawn on demand or transferred to third parties by check or wire. National retail chains are becoming financial supermarkets – offering credit cards, various mutual funds, instalment lending, and insurance along with their vast menus of consumer goods and services; in effect, they would like to become full-service financial intermediaries. At the same time, the traditional intermediaries are moving, as fast as they can obtain government permission, into lines of business from which they have been excluded. Only time will tell how these commercial and political conflicts are resolved and how the financial system will be reshaped (Economic Report of the President 1985, ch. 5).

Portfolio Behaviour of Financial Intermediaries

A large literature has attempted to estimate econometrically the choices of assets and liabilities by financial intermediaries, their relationships to open market interest rates and to other variables exogenous to them. Models of the portfolio behaviour of the various species of financial intermediary also involve estimation of the supplies of funds to them, and the demands for credit, from other sectors of the economy, particularly households and nonfinancial businesses. Recent research is presented in Dewald and Friedman (1980).

Difficult econometric problems arise in using time series for these purposes because of regime changes. For example, when deposit interest rate ceilings are effective, financial intermediaries are quantity-takers in the deposit markets; when the ceilings are non-constraining or non-existent, both the interest rates and the quantities are determined jointly by the schedules of supplies of deposits by the public and of demands for them by the financial intermediary. Similar problems arise in credit markets where interest rates, even though unregulated, are administered by financial intermediaries themselves and move sluggishly. The prime commercial loan rate is one case; mortgage rates in various periods are another. In these cases and others, the markets are not cleared at the established rates. Either the financial intermediary or the borrowers are quantity-takers, or perhaps both in some proportions. Changes in the rates follow, dependent on the amount of excess demand or supply. These problems of modeling and econometric estimation are discussed in papers in the reference above. The seminal paper is Modigliani and Jaffee (1969).

See Also

- ▶ Capital, Credit and Money Markets
- ▶ Central Banking
- ▶ Disintermediation
- ▶ Finance
- ▶ Liquidity
- ▶ Money Supply

Bibliography

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82(6): 1095–1117.

Dewald, W.G., and B.M. Friedman. 1980. Financial market behavior, capital formation, and economic performance. (A conference supported by the National Science Foundation.) *Journal of Money, Credit and Banking*, Special Issue 12(2).

Diamond, D.W., and P.H. Dybvig. 1983. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91(3): 401–419.

Economic Report of the President. 1985. Washington, DC: Government Printing Office.

Federal Reserve System, Board of Governors. 1984. *Balance sheets for the US economy* 1945–83. Washington, DC.

Fisher, I. 1930. *The theory of interest*. New York: Macmillan.

Goldsmith, R.W. 1969. *Financial structure and development*. New Haven: Yale University Press.

Goldsmith, R.W. 1985. *Comparative national balance sheets: A study of twenty countries, 1688–1978*. Chicago: University of Chicago Press.
Gurley, J.G., and E.S. Shaw. 1960. *Money in a theory of finance*. Washington, DC: Brookings Institution.
Modigliani, F., and M.H. Miller. 1958. The cost of capital, corporation finance and the theory of investment. *American Economic Review* 48(3): 261–297.
Modigliani, F., and D.M. Jaffee. 1969. A theory and test of credit rationing. *American Economic Review* 59(5): 850–872.
Schumpeter, J.A. 1911. *The theory of economic development*. Trans. from the German by R. Opie. Cambridge, MA: Harvard University Press, 1934.
Tobin, J. 1963. Commercial banks as creators of 'money'. In *Banking and monetary studies*, ed. D. Carson. Homewood: Richard D. Irwin.
Tobin, J. 1980. *Asset accumulation and economic activity*. Oxford: Blackwell.
Tobin, J. 1983. Financial structure and monetary rules. *Kredit und Kapital* 16(2): 155–171.
Tobin, J. 1984. On the efficiency of the financial system. *Lloyds Bank Review* 153: 1–15.

Financial Intermediation

J. H. Boyd

Abstract

This article deals with the process of financial intermediation: that is, savings and investment flows that are intermediated through organizations such as banks and insurance companies. There are five major topics: stylized facts about financial intermediary organizations and markets; the history of thought about financial intermediation; the theory of financial intermediaries, with an aside on equilibrium credit rationing; the regulation of financial intermediation; and trends in recent research and open research questions.

Keywords

Asset transformation; Bank runs; Banking crises; Banks; Corporations; Default risk; Deflation; Delegated monitoring; Deposit insurance; Discount window; Equilibrium credit rationing; *Ex post* monitoring; Financial intermediaries; Financial; Intermediation; Financial markets; Fractional reserve banking; Friedman rule; Great; Depression; Information economics; Interest rate risk; Liquidity; Monetary policy; Payment services; Private information; Safety net (financial); Savings and loan industry

JEL Classifications

G2

Preliminaries and Introduction

Writing an article such as this requires making some tough decisions about what to include and what not. Many deserving topics in financial intermediation have not been mentioned at all and I cannot begin to cite all the good papers that deserve reference. Primarily, I rely on two excellent survey articles, one that focuses on theory (Gorton and Winton 2003), and another that focuses on empirics (Levine 2005). I received helpful comments from Doug Diamond, Jack Kareken, Ross Levine and Ed Prescott; however, they are totally absolved from any errors that remain.

It is the convention to distinguish between 'financial markets' and 'financial intermediaries'. A financial market is a market in which investors acquire direct claims against ultimate borrowers, usually in the form of debt or equity. A financial intermediary (FI) is a firm that substitutes its own liability for that of some ultimate borrower. That is, an investor lends to the FI and, in turn, the FI lends to an ultimate borrower. I adopt this standard convention even though the distinction is often imprecise. (For example, debt and equity claims are rarely traded directly between the ultimate claimants. Even these are 'intermediated'.) Next, let us turn to some facts about FIs.

The assets of FIs are almost exclusively financial claims. FIs do not have many physical assets, except buildings and computers, and they produce no physical products; thus they are service firms. Important and easily recognizable examples of FIs would include commercial banks, savings and

loan associations, credit unions, life insurance firms, property and casualty insurers, consumer finance companies, and mortgage bankers.

Banks Largest

Commercial banks (hereafter *banks*) are the most important class of FIs, and this has been true for centuries. In developing economies, banks often play a dominant role and may be, essentially, 'the only game in town'. Even in the United States, with its highly developed financial markets, banks accounted for about 14.2 per cent of financial intermediary assets, which is the largest private share, followed by mutual funds at 12.4 per cent (Board of Governors of the Federal Reserve System 2005). This size factor helps explain why banks have been the most-studied class of FI by a wide margin. Banks are also especially important and heavily studied because they create money and thus are the conduit for monetary policy. This article follows the norm and devotes a disproportionate amount of its attention to banks.

Heavily Regulated

FIs are heavily regulated relative to non-financial firms. Most of this regulation is advertised to promote 'safety and soundness', meaning that its stated intent is to reduce the frequency of failures and other problems in the industry. There are four basic forms of regulation: minimum capital requirements, examination by regulatory authorities, portfolio restrictions on asset holdings, and restrictions on who can own or manage an FI. In many countries, there has been a trend towards less intensive regulation of FIs since the mid-1990s, but in these four forms regulation remains obtrusive relative to most industries.

A Large Industry

The FI industry is relatively large. Especially in developed economies, the FI sector is a significant part of the economy, with a substantial share of measured output. In the United States, for example, the total value-added of financial intermediaries (essentially profits, wages and salaries) amounts to about 8.1 per cent of GDP. This makes the US FI sector much larger than (say) the agricultural sector, whose share of total value-added is about one per cent (Bureau of Economic Analysis 2006). Across countries, there is a strong correlation between size and quality of the FI sector and the level of economic development. This relationship is an important topic in development economics but such issues are not considered here. (See financial structure and economic development.)

Organizational Form

In most countries the dominant form of organization for FIs is the corporation; however, there are important exceptions. In particular, many FIs are organized as 'mutuals' or 'cooperatives'. With this alternative form of organization, there is no separate class of shareholders or equity owners, as would be the case in a corporation. For example, in mutual life insurance companies the policy holders are also the owners. In mutual savings and loan associations, the depositors are the owners. These alternative organizational forms are common in the United States, Europe and many other parts of the world.

Recent Trends

Since the mid-1990s, the FI sector has experienced substantial change. The main trends worldwide are towards consolidation (a smaller number of larger firms), diversification (a larger set of financial activities or 'products' offered at the same FI), and internationalization (operating across borders). Almost every part of the world has participated in these developments, excepting sub-Saharan Africa (De Nicolò et al. 2004).

History of Thought on Financial Intermediation

In the 1960s and 1970s, the economic analysis of *FI*s was largely focused on banks, and these were viewed essentially as 'black box' organizations that turned highpowered money (bank reserves) into money. At that time, most intellectual interest in banks derived from their role in creating money, and their being the conduits for monetary policy. In some sense, the study of banking was in those decades incidental to the study of monetary policy

and macroeconomics. There had been an earlier literature on FIs that showed great depth of understanding, but in a nonmathematical, descriptive context. Scholars such as Bagehot, Goldsmith and Schumpeter wrote about, and clearly understood, information asymmetries, liquidity, and so forth. When ambitious scholars, such as Tobin (1969) or McKinnon (1973), tried to incorporate FIs into Keynesian models before the profession had invented the mathematical tools to formally model information and liquidity, the crucial intuitive insights about the role of FIs were absent from the models. Thus, finance became money, and money was simply a stock associated with real capital.

In the mid-1980s a new body of thought emerged and was largely attributable to the seminal work of Diamond (1984) and Diamond and Dybvig (1986). Other significant papers at about that same time included Williamson (1986) and Boyd and Prescott (1986). This new approach to studying financial intermediation stressed that FIs are firms that produce valuable economic services of a variety of kinds, and *explicitly modelled the nature of those services*. This literature was careful to model the profit, share price, or utility-maximizing behaviour of FIs subject to appropriate constraints, and much of this work was done in general equilibrium. More importantly, almost all this work and the large literature that followed featured environments with private information – private in the sense that different agents were endowed with different knowledge. This was a major deviation from the previously studied world of Arrow–Debreu, in which markets are frictionless and perfectly competitive, and all relevant information is common knowledge. It was a critical innovation because in the environment of Arrow–Debreu FIs are irrelevant (cannot increase welfare). In that world FIs are just not very interesting to study in a serious way, and they weren't.

Sequence was also very important in the development of the modern FI literature. Since the post-1983 FI literature almost exclusively employed models with private information, this meant that development of the literature depended on, and naturally followed, advances in information economics thanks to the pioneering work of Akerlof, Hurwitz, Stigler and others. Most likely, this is why earlier efforts to force FIs into Keynesian macro models were a failure; the required tools simply had not yet been invented.

In the next section, I briefly review some of the modern FI models developed in the 1980s and subsequently. Later, in Section 5, I discuss some areas in financial intermediation where, in my judgment, there remain important gaps in our knowledge.

The Theory of Financial Intermediation

Banks and other FIs are firms that take in funds (FI liabilities) through a hypothetical front door, and put out funds (FI assets) through a hypothetical back door. They produce no physical products. To survive, they must earn a profit, meaning that the average rate of return on their assets must exceed the average cost of their liabilities. This spread between asset returns and liability costs must be large enough to cover operating costs (primarily wages and salaries), and to earn a rate of return to equity investors. That FIs earn such positive profits has always troubled critics of the industry (of which there have always been many), who may conclude that FIs are somehow exploiting consumers or businesses. In fact, FIs are permitted to earn these positive interest rate spreads because they provide valuable economic services to the economy, and it is costly to provide these valuable economic services. Let us next consider these services.

One important function, offered by banks but not other FIs, is payment services. This is the 'creators of money' banking function that the old literature stressed, virtually to the exclusion of other FI functions. When we need to execute transactions, we use cash and coin, paper checks, credit cards, and wire transfers. All of these transaction tools are generally provided by banks and for obvious reasons they are economically important.

Another important function of FIs is that they are 'brokers' in the sense that they bring together large numbers of ultimate borrowers and lenders. When they bring these groups together, FIs

substitute their own liabilities for those of ultimate borrowers, and this is what ultimately distinguishes FIs from financial markets. This process has been given many names in the literature ('asset transformation' is common) and understanding it is key to understanding what FIs actually do. Hypothetically, consider one single bank depositor, a wealthy individual, and one single bank borrower, a small business. The bank depositor might have lent directly to the small business through the stock or bond market. Instead, by assumption, he or she lends to the bank in the form of a deposit. In turn, by assumption, the small business borrows from the bank in the form of a commercial loan. The bank places itself in the middle of the exchange and becomes the counter-party to the others.

Why is this valuable? The answer is that bank liabilities typically have different attributes from ultimate borrower liability attributes, ones that are crafted to be desirable to the bank liability holders. If they are made better off, they are willing to lend at a lower rate than they would have required to lend directly. Thus, this process of asset transformation can, and usually does, make both borrowers and lenders better off.

For banks, the general direction of such asset transformation is well understood: bank liabilities will typically have shorter maturity than bank assets, and will be more liquid and less risky. As will become apparent, a key ingredient to this process is that the banks borrow from a large number of creditors, and lends to a large number of borrowers.

Shorter Maturity

Bank liabilities often have shorter average maturity (or duration) than bank assets, and *ceteris paribus* this may make bank liabilities relatively more attractive to savers. Such maturity mismatching exposes banks to an interest rate lottery and the risk that interest rates will increase, in which case they will suffer capital losses. Bank creditors are partially protected against interest rate risk by the bank's equity, at least until that is exhausted. The degree of interest rate risk exposure naturally depends on the magnitude of the asset–liability maturity mismatch, and on how volatile are interest rates. In the 1970s, the US savings and loan (S&L) industry experienced massive losses due to interest rate risk, losses so large as to bankrupt much of the industry as well as its government insurer, the Federal Savings and Loan Insurance Corp (FSLIC). The S&Ls' maturity mismatch was substantial, and interest rates had become extremely volatile by historical standards. However, the savings and loan industry should not be blamed for this sad experience. Government regulations essentially forced this industry to borrow short and lend long.

Since the mid-1990 banks and other FIs have become clever in finding ways to hedge interest rate risk in the forward, futures and swap markets. (Of course, someone still has to bear the aggregate risk.) Also, there is some evidence that, in the United States at least, FIs have in recent years become less willing to expose themselves to interest rate risk. As a practical matter, however, it is difficult to accurately measure the maturity mismatch of banks, and standard duration methods may not work very well for this industry. That's because a substantial proportion of bank liabilities are in the form of demand (checking) deposits. For these liabilities, the technical maturity is instantaneous but the true maturity is much longer, depends on economic conditions, and must be empirically estimated.

More Liquid

Bank liabilities, especially deposits, are more liquid than bank assets. This is another desirable form of asset transformation since, *ceteris paribus*, lenders like to hold liquid assets. The liquidity provision function has been heavily studied by scholars, and the seminal reference on the topic is Diamond and Dybvig (1986). Now, liquidity is hard to define, let alone understand, and it may help to consider a simple theoretical environment, similar in some ways to the more complicated environment studied by Diamond and Dybvig (1986). Imagine a world in which there are only two assets: gold coins and land. By assumption, gold coins are perfectly liquid and can be spent at any time but earn no rate of return. Land, on the other hand, is highly

productive but illiquid. It is hard to sell land in an emergency, and possibly it can't be sold at all. All agents in this economy have a known, say one per cent, chance of an 'emergency', the occurrence of which is independent across agents. In an emergency, agents desperately want to have all their wealth immediately so they can consume it. Now, consider the problem facing individual agents. If they put all their wealth in coins (land) they will do well 1 (99) per cent of the time; however, they will do very badly 99 (1) percent of the time. Common sense suggests that the best strategy will be to split up their holdings, and if you guessed that you would be right at least for most preferences. Even then, however, agents are not doing as well as they potentially could in either state of the world.

Next, assume a bank is organized, which offers each individual a deposit account that can be *redeemed in gold coins on demand*. Further, assume the bank puts 1 per cent of its assets in gold coins and 99 per cent in land. Now, if the bank deals with a sufficiently large number of depositors, it will have enough coins to just cover withdrawals and all the remaining can be invested in highly productive land. Everyone is better off than they could have done on their own account.

This kind of an arrangement is usually referred to as 'fractional reserve banking'. The key to its smashing success is diversification across a large number of depositors, and the fact that depositor withdrawal demands are independent. Now, as Diamond and Dybvig are quick to point out, this idealized solution may not always work out in practice. Suppose, for example, that emergency withdrawals become correlated, perhaps because there is a war. Then the bank can easily run out of coins, fail on its obligations, and land must be inefficiently liquidated. Even worse, just a false rumour of war could send too many depositors to the bank and cause it to fail. This sort of occurrence is called a 'bank run' and these have been quite common both historically and in recent times. In an imperfect world where withdrawals may be correlated and bank runs are possible, every bank faces a fundamental and unpleasant trade-off: if it holds a high fraction of gold coins (reserves) risk of insolvency will be low, but the average rate of return on its assets will be low. If it holds a high fraction of land (earning assets) its average rate of return on assets will be high, but its risk of insolvency will be high. There is a large literature on this topic, much of which is referenced in Gorton and Winton (2003).

Less Risky

Bank liabilities are on average less risky than bank assets, and obviously this tends to make bank liabilities *ceteris paribus* more attractive. Now, bank liabilities can be less risky than the representative bank loan for a variety of reasons. One is that banks often place some fraction of their assets in default-risk-free government securities. A second reason is that banks raise part of their funds in the form of equity, and the bank's shareholders must suffer a total loss before liability holders lose. A third reason is that banks hold portfolios of different kinds of loans that are diversified by industry and geography, so that their loan portfolio is less risky than its individual components. A fourth reason is that in most countries bank deposits are fully or partially insured by government.

In addition, banks are very good at determining to whom to lend, and in setting loan terms for those who are funded. This topic has been heavily studied in the FI literature and the reader can find many studies under the headings 'adverse selection', 'sorting' and 'screening' in Gorton and Winton (2003). In most of these models, some loan applicants are better credit risks than others, applicants know their own types, and are willing to misrepresent (say they're good when they're bad). FIs do not know the applicants' types, although it is conventional to assume that everyone knows the underlying distribution of applicant types. The FI's objective is to accept (reject) good (bad) applicants where possible. In some but not in all cases, it is possible to adroitly choose terms of lending such that good applicants voluntarily sign up, and bad applicants withdraw. In other cases, the best strategy is simply to accept (reject) all applicants.

Another important aspect of lending, and an aspect at which FIs excel, is monitoring borrowers

after they have received the money. This '*ex post* monitoring' has also been heavily studied in the FI literature. Once they have the money, borrowers may take actions that reduce their probability of repaying, or events beyond their control may have the same effect. To protect their interests lenders normally pre-specify loan covenants that state what happens in such cases, and they monitor borrowers to enforce these covenants. An example that homeowners will understand is a residential mortgage: to protect its interests, the lender must be sure that property taxes are being paid, and that the house is fully insured. Now, it is often the case that loans are large relative to the wealth of individual agents in the economy. This naturally occurs because many production technologies exhibit economies of scale. For example, an automobile plant must be of a particular size to be efficient, and few if any agents can fund such an investment with their own wealth. Therefore, to fund a loan often requires obtaining financing from several agents simultaneously. Unless FIs are present there is a coordination problem among the several lenders, and it is a problem first studied by Diamond (1984) and Williamson (1986).

Monitoring of borrowers is costly, and no one wants to do it if they don't have to. Now, for simplicity, assume that there are just two lenders for a given loan, lender A and lender B. Now, A (B) may assume that B (A) will monitor, in which case neither lender actually does. This is obviously undesirable because their interests are not being protected. Alternatively, lender A and lender B might both be conservative, assume the other is unreliable, and monitor themselves. In that case there would be redundant monitoring which is unnecessary and wasteful. Clearly, what is needed is an arrangement in which all lenders agree to have *ex post* monitoring done by a single, efficient 'delegated monitor'. What is critical, if such an arrangement is to work, is that the delegated monitor finds it in its own interests (incentive compatible) to actually do the work as promised. Otherwise, it might be necessary to monitor the monitor, which obviously would be inefficient, too. Diamond (1984) and Williamson (1986) showed that efficient *ex post* monitoring can be achieve by a bank that pools funds from many depositors and uses the proceeds to make many loans.

Summary

In a world in which different agents have different information sets FIs earn a positive interest spread between their average asset returns and average liability costs, in return for providing valuable services. They are brokers between ultimate borrowers and ultimate lenders, and they provide payments services. They transform ultimate financial claims in the sense that their liabilities have different attributes from their assets. Typically, their liabilities are shorter in maturity, more liquid and less risky; thus, such liabilities are more desirable to savers. This process of 'asset transformation' is not without risk. FIs are exposed to interest rate risk and particularly vulnerable to unexpected interest rate increases. We discussed the case of the US savings and loan industry and its devastating exposure to interest rate increases. Due to their liquidity provision, banks are exposed to the risk of bank runs. Bank runs have been common historically, and still have occurred with some frequency in the modern wave of banking crises. Finally, all FIs are exposed to default risk when their loans or other investments do not pay off in a timely manner.

An Aside on Equilibrium Credit Rationing

When economists began studying intermediation environments with private information, in which agents could withhold the facts, intentionally deceive one another, and so on, all manner of new and interesting results were obtained. One seminal model of financial intermediation featured an outcome called 'equilibrium credit rationing' (Stiglitz and Weiss 1981). In such cases, at the equilibrium rate of interest there is excess demand in the sense that some would-be borrowers are denied access to credit. This is quite at odds with a classical market equilibrium, and immediately raises the question, 'why don't lenders just increase the rate of interest to a level at

which demand equals supply?' A variety of answers to this question can be found in the literature, reflecting the different environments that have been shown to produce equilibrium credit rationing. For one example, assume that credit applicants are of two types, good and bad, and that lenders take account of borrower heterogeneity in their rate setting. Then, it can be the case that for sufficiently low interest rates both good and bad will borrow, but above some threshold rate r^* good types become unwilling to borrow. In such cases lenders may find it optimal to set the rate at r^* even though there is excess demand at that rate. A second example is an environment with moral hazard in the form of a bad action that borrowers may take *ex post* (such as increasing the risk of their investment project). For some parameterizations, when rates are below a threshold r^+, borrowers will not take the bad action, but above r^+ they will. As in the case above, it may be optimal for lenders to set the rate at r^+, thus avoiding the bad action, and resorting to credit rationing.

These first two environments are with private information: however, a third one can result in equilibrium credit rationing even when all information is public. Imagine that default by borrowers results in a deadweight loss – for example, an out-of-pocket bankruptcy cost. Then, the probability of costly default directly depends on the rate of interest, and the higher that rate is the higher the default probability is. Increasing the rate of interest increases the expected rate of return to lenders in good (non-default) states, but also increases the probability of default which is costly to both parties. Depending on the distribution of possible returns facing borrowers, it may be that raising the rate beyond some threshold r^- is futile in the sense that the marginal cost exceeds the marginal benefit. In these cases, rates above r^- are harmful to both parties and will never be observed in equilibrium. Yet it may also be true that *r*-plus is too low to clear the market, and equilibrium credit rationing will again be observed (Williamson 1986).

Arguably, equilibrium credit rationing is a topic where theory leads measurement. There has not been a lot of good empirical work on credit rationing per se, primarily because it is so hard to do right. Credit rationing equilibria are off the usual demand and supply curves that econometricians like to estimate, and they may exhibit nasty jumps, discontinuities, and so on. If the theorists are right, however, and credit rationing is popping up all over, more empirical work would be useful, especially in the area of finance and development.

Regulation

Banks and other FIs are, almost without exception, rather heavily regulated. This is true in virtually all countries and has been true for centuries. There are at least three reasons for this special and obtrusive regulatory treatment. First, banks are the conduit for monetary policy, and problems in banking are likely to interfere with monetary policy conduct. Second, it is widely believed that bank failures may result in negative externalities (social costs). And third, governments may find it irresistible to control a critical industry that creates money and allocates a large fraction of investment capital. Some recent work has emphasized the importance of political economy issues for regulation, in particular arguing that it is unlikely that bank regulation can contribute positively to social welfare in economies with weak and/or corrupt governments (Barth et al. 2006).

The Great Depression was a difficult time for banks in the United States and many other countries, and during the late 1920s and early 1930s there were literally thousands of bank failures worldwide. Many of these were associated with bank runs and panics. In response, many nations substantially beefed up their regulation of FIs and put in mechanisms such as deposit insurance to reduce or eliminate the prevalence of bank runs. For example, the Federal Deposit Insurance Corporation was created by US federal legislation in 1933. Beginning in the mid-1930s, the industry stabilized (at least in developed nations), and went through a period of relative calm that lasted for about three decades. Many observers believed that these policy interventions had solved the problem of instability in banking; but that was not to be. Beginning in roughly the mid-1960s, a new

wave of banking crises affected well over 100 nations. Banking crises – some of them severe – have been recently experienced in developing and developed economies alike.

No one knows for sure what has caused this interesting historical sequence of events in banking, but many scholars have emphasized that *policy interventions intended to stabilize the industry may have actually had opposite effect*. In most countries, banks have access to emergency borrowing from the government (a Discount Window), and have some form of government insurance to protect depositors. Additionally, there is a common practice known as 'too big to fail' whereby governments will prop up their very largest FIs if they get into trouble. This package of interventions is widely referred to as 'the safety net', and it has been very heavily studied. Most of the literature on this topic concludes that, whatever the benefits of a safety net, it also distorts bank incentives in a perverse way. Depositors and other bank creditors don't care how much risk the bank takes (they are protected by government), and normal market risk-constraining mechanisms become ineffective.

In the presence of a safety net, banks may have an incentive to take on more risk *ceteris paribus* than otherwise; indeed, they may even become risk lovers who intentionally seek out investments with low expected returns and high variance. It's not hard to see why this is so. If an FI has very risky investments and these payoff, all the profits go to FI shareholders. If they don't payoff the FI goes broke, but the resulting losses are mostly absorbed by government. In essence, this is a 'heads I win tails you lose' gamble. Perhaps the most dramatic evidence of this distortion turned up during the U.S. S&L crisis. At that time, many S&Ls were obviously bankrupt but could not be closed down since their federal deposit insurer, the FSLIC, had run out of money. Many such institutions gambled for redemption by taking extreme risks. If they were lucky enough they might survive, and if not…well, they were already broke.

As of 2007, solving the problems associated with the safety net is arguably the most vexing policy issue facing FI regulators and scholars of that industry. Many regulatory interventions, such as restrictions on asset holdings, attempt to control FIs' behaviour but do not deal with the fundamental distortion of risk incentives. Other regulatory interventions such as capital regulation are intended to reduce FIs' distortion of risk incentives, but may not be effective (Hellmann et al. 2000). FIs have a natural tendency to try to get around all these regulations, pursuing strategies that render the regulations ineffective. On the other hand, getting rid of the safety net would have its own risks, and it is far from obvious how governments could ever credibly commit to a policy of no FI bailouts. This issue is probably best described as important but unfinished business.

Trends in Recent Research, and Open Research Questions

1. As discussed earlier, the modelling of financial intermediaries has come a long way since the mid-1980s, and most modern macroeconomic models reflect that reality. Even so, there is still recent work that reflects old ways of thinking about FIs. To make this point I provide just one example: the ongoing discussions of the so-called 'Friedman Rule'. This rule, in simplest form, calls for a monetary policy that produces a rate of *deflation* such that the real rate of return on bank reserves equals the rate of interest on real investment. Then, it is argued, banks will voluntarily hold all their assets in the form of reserves, and bank runs, crises, and so on will never happen. Bruce Smith (whose death in 2003 was a great loss to economics) makes it beautifully clear that this oncebeguiling idea should be relegated to the history of economic thought (Smith 2002). Application of the Friedman rule may indeed result in risk-free banks. However, except for the provision of payments services, it precludes banks from making any of their valuable economic contributions detailed by Diamond (1984); Diamond and Dybvig (1986) and others, and as discussed earlier.
2. Boyd and Prescott (1986) have a theorem that financial intermediary coalitions composed of

large numbers of agents can support allocations that cannot be supported with decentralized markets, and are efficient subject to resource and incentive constraints. As lamented by Green and Zhou (2001), virtually all subsequent theoretical research on FIs has studied decentralized (market) environments. Now, this could be just a matter of preferences amongst theorists as to the most interesting and tractable environment in which to study FIs. It's not, in my opinion, and this topic is of more than theoretical interest. Boyd–Prescott financial intermediary coalitions look (at some high level of abstraction) like mutual or cooperative FIs. It is fact that over several continents and many centuries mutual FIs seem to endogenously spring up with great regularity. When a class of arrangements is 'revealed preferred' so often, there is probably a good reason for it. There has been some theoretical research on this topic, but arguably not enough.

3. Virtually all of our general equilibrium models with FIs force agents into discrete silos: for example, an agent must choose to become a producer (borrower), a consumer (lender), or an FI. In reality we often observe organizations that are both producers and financial intermediaries at the same time (for example, General Electric or Cargill). Moreover, we sometimes see firms radically change their blend of activities. For example, in a few years Enron evolved from a production firm to a financial intermediary. I am aware of only one study (Bhanot and Mello 2006) that allows, in a serious way, for endogenous choice of FI *and* non-FI activities in the same organization. More work along these lines could be useful.
4. As discussed, banks, even very simple ones, perform a number of economic functions *simultaneously*: brokerage, payments service provision, maturity transformation, liquidity provision, and default risk reduction. This is what we observe in reality and there is undoubtedly a reason. Yet our theoretical models tend to isolate these economic functions and look at them one at a time. Only a few studies have seriously looked at the jointness in providing even two services simultaneously (Kasyap et al. 2002). This separation of functions is done for tractability, and even then our models can become complex. Putting all of these features in a model simultaneously becomes technically daunting, but it needs to be done. There are undoubtedly interesting interactions or synergisms among these activities, and we cannot learn about those by studying them individually.

See Also

- ▶ Finance
- ▶ Financial Structure and Economic Development

Bibliography

Barth, J., G. Caprio, and R. Levine. 2006. *Rethinking bank regulation: Till angels Govern*. Cambridge: Cambridge University Press.

Bhanot, K., and A. Mello. 2006. *Should production and trading activities be separated?* Working paper: University of Wisconsin.

Board of Governors of the Federal Reserve System. 2005. *Flow of funds accounts*. Online. Available at http://www.federalreserve.gov/RELEASES/z1/. Accessed 16 Jan 2007.

Boyd, J.H., and E.C. Prescott. 1986. Financial intermediary coalitions. *Journal of Economic Theory* 2: 211–232.

Bureau of Economic Analysis. 2006. *Industry economic accounts*. Online. Available at http://bea.gov/bea/dn2/home/annual_industry.htm. Accessed 3 Feb 2007.

De Nicolò, G., P. Bartholomew, J. Zaman, and M. Zephirin. 2004. Bank consolidation, internationalization and conglomeration: Trends and implications for financial risk. *Financial Markets, Institutions & Instruments* 13(4): 173–217.

Diamond, D. 1984. Financial intermediation and delegated monitoring. *Review of Economic Studies* 51: 393–414.

Diamond, D., and P. Dybvig. 1986. Banking theory, deposit insurance, and bank regulation. *Journal of Business* 59: 55–68.

Gorton, G., and A. Winton. 2003. Financial intermediation. In *Handbooks in the economics of finance, volume 1A: Corporate finance*, ed. G. Constantinides, M. Harris, and R. Stulz. Amsterdam: North-Holland.

Green, E., and R. Zhou. 2001. Financial intermediation regime and efficiency in a Boyd–Prescott economy. *Carnegie-Rochester Series on Public Policy* 54: 117–129.

Hellmann, T., K. Murdoch, and J. Stiglitz. 2000. Liberalization, moral hazard in banking and prudential

regulation: Are capital requirements enough? *American Economic Review* 90: 147–165.
Kasyap, A., R. Rajan, and J. Stein. 2002. Banks as liquidity providers: An explanation of the coexistence of lending and deposit-taking. *Journal of Finance* 57: 33–74.
Levine, R. 2005. Finance and growth: Theory and evidence. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.
McKinnon, R. 1973. *Money and capital in economic development*. Washington, DC: Brookings Institution.
Smith, B.D. 2002. Monetary policy, banking crises and the Friedman rule. *American Economic Review* 92: 128–134.
Stiglitz, J., and A. Weiss. 1981. Credit rationing in markets with imperfect information. *American Economic Review* 71: 393–410.
Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit, and Banking* 2: 461–472.
Williamson, S.D. 1986. Costly monitoring, financial intermediation, and equilibrium credit rationing. *Journal of Monetary Economics* 18: 159–179.

Financial Journalism

Richard Fry

Financial journalists write in the press, or talk on radio or television, about the financial markets and all the factors that move them. The term 'financial' to describe this work is traditionally used in England, but the more senior writers cover a much wider ground. They report and discuss any matter that may be of professional interest to the alert businessman. Their field includes economic analysis and comment. Their personal views on events are often expressed with a freedom rarely allowed to other journalists, and they are sometimes influential both in moving markets and in shifting public and political opinion.

There are two reasons for this state of affairs in England, one technical and one historical. The technical reason is that British Ministers and senior officials, as well as leaders of business and banking, prefer to brief journalists 'off the record' and anonymously. This is also true in France but not generally in the USA or in West Germany. The British journalist, therefore, expresses as his own view something he may have learnt from an 'inside' source, and it is often difficult for the reader of the British financial press to spot the line between official thinking and the writer's own opinion. The system has many critics, but it accounts in part for the unique role of financial journalists in England.

Historically this status goes back to the 19th century, when a number of outstanding personalities were drawn into financial and economic journalism. After the Napoleonic wars, London gradually became the chief financial centre of Europe, taking over from Amsterdam. During the industrial revolution a large, wealthy middle class had emerged which had surplus savings and looked for investment opportunities. The growing demand for financial information, analysis and advice was met by newspapers and weekly journals on the basis of advertising by company promoters, banks and (later) joint-stock companies. The London *Times* appointed its first financial editor, Thomas Massa Alsager, in 1817. He was not an economist but a businessman with wide cultural interests. He opened an office close to the stock exchange and the Bank of England, wrote the daily financial article and organized the collection of 'mercantile and foreign news'. He became a friend of the Rothschilds, made a fortune, and did not hesitate to criticize the Bank of England. For some years he stood alone in warning investors that the great boom in railway promotions was bound to collapse. The *Times* lost a great deal of advertising revenue but the proprietors were high-minded and Alsager was proved right.

Another milestone was the founding of *The Economist* in 1843 as 'a political, literary and general newspaper' by James Wilson, a banker and Member of Parliament who had been Financial Secretary to the Treasury and later became Finance Member of the Council of India. Wilson started the weekly paper, which he and his family owned, to spread the ideas of free trade, free enterprise and political reform. From the start it had a substantial statistical section, and soon monetary and banking subjects became prominent. James Wilson wrote several articles in every issue, including the important one on the 'Money Market'. In 1857 he engaged Walter

Bagehot, a rising economist and banker, to write for him on banking. Bagehot soon widened his subject; he also married Wilson's daughter, and in 1859 when Wilson went to India he became sole editor and remained so until his death in 1877. Bagehot gained a high reputation as a financial expert, adviser to governments and author of books. (Keynes said Bagehot wrote *Lombard Street* in 1873 in order to 'knock two or three fundamental truths into the heads of City magnates'.) Throughout, week after week, Bagehot remained a journalist and a crusader for reform. For example, he warned of the danger of allowing the pound sterling to be used as an international currency. When that became inevitable because for a time the pound was the only currency freely convertible into gold, he demanded that the Bank of England should build up a separate gold reserve so that withdrawals of foreign deposits should not deflate the British economy. That theme has remained alive for more than 100 years.

The blend of political, economic and financial subjects has been a successful formula for *The Economist* ever since. It has also been a model for some other papers that came a little later. The first daily newspaper devoted to financial and business matters was the *Financial News*, founded in 1884, followed by the *Financial Times* in 1888. The two papers had periods of success and weakness. They merged in 1945 with the title *Financial Times* and the combined paper has greatly widened its scope and increased its circulation.

A great revival of financial journalism took place in England after World War I, when a number of gifted young university graduates were recruited, first by the *Manchester Guardian* and *The Economist* and a few years later by the *Financial News*. This group spread quickly to other publications, and though it lost many of its young members to high positions in government service, banks and universities, it raised the profession to a status not equalled in any other country. The opening was provided in the early 1920s by a surge of public demand for information and comment on a bewildering series of events. War debts, reparations, the destruction of currencies by inflation, the 1925 restoration of the gold standard, mass unemployment, the Wall Street crash of 1929 and the world-wide depression that followed, the 1931 sterling crisis – all these required some expert knowledge for proper understanding and discussion. The new type of financial journalist was picked and trained to meet this demand.

The new wave was started by Oscar Hobson (1886–1961). He went to King's College, Cambridge, partly at the same time as Maynard Keynes, whose views he sternly opposed all his life. After taking first class degrees in classics and mathematics and a brief spell in a London bank, Hobson became financial editor of the *Manchester Guardian*, where he took an active part in the public arguments of the 1920s. In 1929 he was made editor of the *Financial News*, which had just passed into new and ambitious hands. There he found the first few of the new type of financial journalists already in place, and he added quickly to them, assembling a brilliant group that became the chief nursery for British financial journalism. In 1934, after a sharp dispute over policy with his publisher (Brendan Bracken), Hobson left and became 'City Editor' of a general daily newspaper, the liberal *News-Chronicle*, where he was given less than half a page to cover financial news and comment. He managed to write each day a decisive little essay in a few square inches. One day he was summoned to the Governor of the Bank of England, Montagu Norman, who asked why he did not collect his daily essays to make a book. Hobson laughed. The Governor opened his desk, took out a pile of clippings with Hobson's daily writings and said: 'I have the book ready here, and I have arranged a publisher for it.' The book was published and went into many editions.

Hobson always kept close touch with academic economists. He was a member of the Council of the Royal Economic Society, a governor of the London School of Economics, and was made a knight. The economist Lionel Robbins wrote in his autobiography that Hobson was 'one of the creators of modern standards of professional excellence in financial journalism'.

Among these 'creators' one must certainly mention Sir Walter Layton (later Lord Layton) who arrived at *The Economist* in 1922. The

editor of that paper, once appointed, is given very wide independence. Layton used his to introduce new men and new ideas. He formed a strong group of editorial writers, many of whom later left to become editors of other papers. He added greatly to the statistical and business coverage of *The Economist*. When he left in 1938 he had restored *The Economist* as a potent influence in British public life. This work was interrupted by World War II but resumed after it by Geoffrey Crowther, whose editorship gained the paper its important international readership. Crowther, too, ended his career as Lord Crowther, and he left a successful and respected enterprise behind him.

Since World War II British financial journalism has maintained its standing. Some of the leading writers have joined the profession as very young men – and more recently women – and have made their reputations. Like their predecessors they were helped by the fact that Britain is highly centralized. In contrast to the United States and West Germany (and Italy), where the political capital is separate from the financial one, London is both the seat of government and the centre of business and finance. The bank head offices, the stock exchange, the money market, the insurance and shipping markets and many others are located, with many of their professional adjuncts, in the famous 'square mile' of the 'City'. London press, radio and television dominate the country. It is true that France, too, is centralized in Paris; but the French press has not, in the past, enjoyed sufficient independence from either its proprietors or the government to build up a reputation for financial and economic authority, though a beginning has been made.

Moreover, in Britain financial journalists tend to come from the same background as people in government. In the 1970s and 1980s several of the leading economic journalists worked for a time in a government department or in the Bank of England. At one time the Treasury organized a confidential seminar for financial journalists to explain its latest forecasting methods. Some financial journalists have become Treasury officials, while Nigel Lawson was a financial journalist for some years before he went into politics and rose to become Chancellor of the Exchequer in Margaret Thatcher's second government.

It is not always necessary to have close contacts with the administration. One of the most influential financial journalists in Britain after World War II was Harold Wincott. He did not attend university but worked as a statistician for a stock exchange firm. In 1930 he joined the *Financial News* as a sub-editor, working mostly at nights. He soon began to write in various sections of the paper and attracted attention. In 1938 he was made editor of the weekly *Investor's Chronicle*, then owned by the Financial News. That was the platform he used for a number of years to comment on the financial markets. Wincott did not trouble about government secrets. He looked at what the government was doing and found it almost lunatic. In his simple, humorous style he pulled to pieces the system of regulations and controls that had been erected to keep business in its place. His weekly commentaries moved to the centre page of the *Financial Times* and became very popular. He was one of a small group of British writers who played a powerful part in restoring a belief in the market economy among the voting public. When Wincott died suddenly in his fifties some friends, led by the economist Lionel Robbins, launched an appeal for funds to form a new foundation for the spreading of these ideas. They received many times as much money as they had expected and duly set up the Wincott Foundation, which now finances a number of research projects, lectures and publications, besides awarding a prize for the 'financial journalist of the year'.

Another writer of influence was George Schwartz, an economist who had come from Vienna to the London School of Economics where he served for many years as a lecturer. He often prepared statistical and other material for J.M. Keynes. He discovered in middle age that he had a powerful gift for writing, and after a few successful attempts he left academic work to write a weekly article on economic life as he saw it for the London *Sunday Times*. Schwartz believed profoundly in the truth of liberal economics and the rightness of allowing people to strive for their own advancement. He wrote in simple terms that

millions could understand, and his views gained much influence. Among his regular readers was the Queen.

Not many professional economists have made a success in financial journalism in Britain (unlike the United States). Economists are, of course, asked from time to time to contribute a comment or forecast to a daily newspaper, and some of the many economists employed by stockbrokers are often quoted for their views on specific situations. But there are no outstanding reputations. John Maynard Keynes had a close relationship with the *Manchester Guardian*. After World War I he agreed to edit for the paper a series of supplements on 'The Reconstruction of Europe'. To each of these twelve supplements (April 1922–August 1923) he made a contribution of his own; much of this material was later incorporated in his *Treatise on Money* (1930) and the *General Theory of Employment, Interest and Money* (1936). In addition Keynes persuaded many of the leading statesmen, bankers and economists of Europe to write for the supplements. They were widely read and translated into a number of languages. Later, Keynes repeatedly launched or tried out new policy ideas in newspaper articles. The brilliance of his writing alone assured these pieces a wide readership, but one could hardly call Keynes a financial journalist.

A few newspapers and journals have been mentioned above to illustrate the curious role of financial journalism in Britain. For a complete picture one would, of course, have to mention many more. 'Financial Editors' (the common description of the chief financial writer, who is usually also in charge of the reporting staff) are as a rule persons with knowledge, experience and ideas, though not all of them have had degrees in economics. The best of them are able to talk on equal terms with bank presidents, high officials and even economists. Many talented young people have been attracted to the financial services industry (banks, investment firms, stockbrokers etc) for at least a generation, and financial journalism has had its share of the recruits. Radio and television have greatly widened the scope. Financial reports form part of the main news programmes, and there are several serious analytical or discussion programmes, mainly run by former newspaper journalists.

While the standing of British financial journalists may be unique, the work itself is, of course, being done in many other countries. In the United States the *Wall Street Journal* has probably as much influence with the US administration as the *Financial Times* has in London. The *New York Journal of Commerce* has a much smaller circulation but contains much material of the highest quality. Among general newspapers one finds thorough, intelligent business sections in the *New York Times, Washington Post, Boston Globe, Christian Science Monitor, Chicago Tribune* and *Los Angeles Times*. Others like the *Miami Herald, Dallas Times Herald, Dallas Morning News* and *St Louis Post Dispatch* might claim inclusion in the list. American magazines specializing in this field include *Fortune*, *Forbes* (both bi-monthly), *Business Week* (weekly) and *Barrons*. Dun's *Business Month and Financial World* might be added as well as *Financier*, a monthly journal which deals seriously with policy issues involving the business and financial community. Several television programmes have gained importance. Having written this very selective list, one is left with the fact that in the United States the journalists whose names are well known and carry a certain glamour are the political commentators, not the financial journalists.

West Germany has a long tradition of good financial journalism, which goes under the broader description *'wirtschaftlich'*. Many publications and their titles have changed since World War II. The chief daily business newspaper is the *Handelsblatt*, published in Düsseldorf, and for stock market subjects the daily *Börsenzeitung*. Most influential is the business (Wirtschafts-) section of the *Frankfurter Allgemeine Zeitung* which runs to 5–6 pages. Its reputation goes back to the pre-war *Frankfurter Zeitung* and it was this famous newspaper which first issued a small book entitled 'How to read the commercial section of a daily newspaper' known to students of journalism for generations. This has now grown to a book of 550 pages and is called *So nutzt man den Wirtschaftsteil einer Tageszeitung*, edited by Jurgen Eick.

Other business and financial information is to be found in the *Wirtschaftswoche* of Düsseldorf, the *Zeitschrift für das gesamte Kreditwesen*, Frankfurt (mainly banking and financial policy) and in a number of general newspapers including *Die Welt* and *Die Zeit*. German financial journalists in the leading positions enjoy a little of the prestige that clings to their British counterparts.

The Swiss *Neue Zürcher Zeitung* occupies a special place in Europe for the quality and responsibility of its financial and economic pages. It is not surprising that Dr Franz Aschinger who for many years edited that section went on to be economic adviser of one of the three big Swiss banks and finally professor at the St Gallen university.

In France, independent financial and economic journalism is relatively new and developing slowly. Daily financial information is supplied mainly by the French-international news agency AGEFI (Agence économique et financière) which issues four editions each weekday of 12–16 pages each. It has an able staff of reporters and strong correspondents in the main financial centres. Another daily publication is *Les Echos* which concentrates on stock market information. Among general daily newspapers *Le Figaro* has an influential finance section and a weekly supplement called 'La vie économique'. The financial section of *Le Monde* and its weekly economic report are also of good quality.

The two leading weekly journals are *La Vie Française*, which now prints 120–140 pages and covers the French business scene, particularly with descriptions of companies, personalities and regions; and *Le Nouvel Economiste*, about 100 pages, covering general business subjects. Appearing twice a month is *L'Expansion*, an impressive magazine modelled in format on the American *Fortune*. It contains serious economic analysis, articles on business personalities and corporations. After a hesitant start its circulation reached 170,000 in 1984. A monthly journal specializing in banking and monetary subjects is the revue *Banque*, published by the association of French banks. It often contains serious papers on economic problems. A financial radio programme made with professional skill forms part of the nightly 'Europe No.1 news'.

In Japan the leading daily business newspaper is *Nippon Kezai Shimbum* which combines financial and corporation news with market reports. Two general daily newspapers have substantial and respected business sections: *Asahi Shimbum* and *Yomiuri Shimbum*. There are two weekly economic journals: *Toyo Kezai Shimposha* (sometimes described as *The Economist* of Japan) and *Weekly Diamond*, published in English and dealing mainly with the investment markets. Financial journalism is a recognized profession in Japan, though its independence and social standing is not quite the same as in the West.

There are, of course, financial publications of high quality in some countries not mentioned above: Italy, India, Singapore, Hong Kong certainly have examples. Financial journalism has become a worldwide occupation.

See Also

- ▶ Bagehot, Walter (1826–1877)
- ▶ Crowther, Geoffrey (1907–1972)
- ▶ Einzig, Paul (1897–1973)
- ▶ Layton, Walter Thomas (1884–1966)
- ▶ Shonfield, Andrew Akiba (1917–1981)
- ▶ Wilson, James (1805–1860)
- ▶ Withers, Hartley (1867–1950)

Financial Liberalization

Romain Rancière, Aaron Tornell and Frank Westermann

Abstract

Financial liberalization has led to financial deepening and higher growth in several countries. However, it has also led to a greater incidence of financial crises. Here, we review the empirical evidence on these dual effects of

financial liberalization across different groups of countries. We then present a conceptual framework that explains why there is a trade-off between growth and incidence of crisis, and helps account for the cross-country difference in the effects of financial liberalization.

Keywords

Bail-out guarantees; Banking crises; Boom–bust cycles; Capital account liberalization; Contract enforceability; Credit growth; Currency crises; Economic growth; Equity market liberalization; Financial liberalization; Financial openness; Foreign direct investment; India; Insolvency risk; International flows; Investment; Investment subsidies; Lending booms; Portfolio flows; Probit models; Prudential regulation; Skewness; Thailand; Tradable and non-tradable Sectors; Trade liberalization

JEL Classification

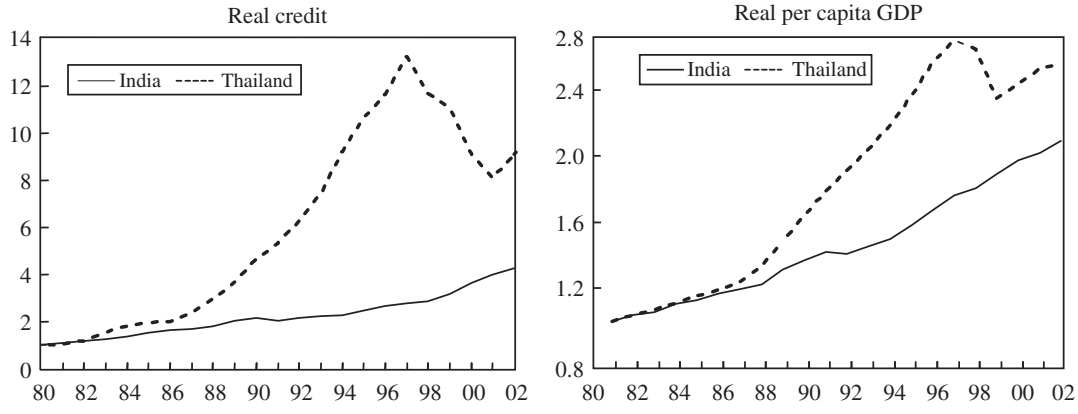
F3; F4

Financial liberalization (FL) refers to the deregulation of domestic financial markets and the liberalization of the capital account. The effects of FL have been a matter of some debate. In one view, it strengthens financial development and contributes to higher long-run growth. In another view, it induces excessive risk-taking, increases macroeconomic volatility and leads to more frequent crises. This article brings together these two opposing views.

The data reveals that FL leads to more rapid economic growth in middle-income countries (MICs), but does not have the same effect in low-income countries (LICs). In MICs this process is not smooth, however: It takes place through booms and busts. Indeed, MICs that have experienced occasional financial crises have grown faster, on average, than non-liberalized countries with stable credit conditions. In LICs liberalization does not lead to higher growth because their financial systems are not sufficiently developed so as to permit significant increases in leverage and financial flows.

The contrasting experiences of Thailand and India illustrate these dual effects. Thailand, a liberalized economy, has experienced lending booms and crises, while India, a non-liberalized economy, has followed a slow but safe growth path (see Fig. 1). In India GDP per capita grew by only 99% between 1980 and 2002, whereas Thailand's GDP per capita grew by 148%, despite a major crisis. As will be shown below in a set of data analyses, this trade-off exists more generally across MICs.

Asymmetric financial opportunities across sectors are key to understanding the effects of FL. In particular, in MICs contract enforceability



Financial Liberalization, Fig. 1 Safe vs. risky growth path: a comparison of India and Thailand, 1980–2002 (*Note*: The values for 1980 are normalized to 1. *Sources*: International Financial Statistics (IMF) and Word Bank Development Indicators)

problems affect the tradable (T) and non-tradable (N) sectors differently. Many T-sector firms are able to overcome these problems and gain access to international capital markets, whereas most N-sector firms are financially constrained and depend on domestic banks for their financing. Trade liberalization promotes faster productivity growth in the Tsector, but is of little direct help to the N-sector. By allowing banks to borrow on international capital markets, FL leads to an increase in investment by financially constrained firms, most of which are in the N-sector. However, while FL increases investment, it also increases borrowers' incentives to take on insolvency risk because there are implicit and explicit bail-out guarantees that cover lenders against systemic defaults. This is why greater leverage and growth is associated with aggregate financial fragility and occasional crises.

In the rest of this article we describe the ways in which FL has been measured and the empirical estimates of its effects on growth and crises. We then present a conceptual framework and a review of the policy issues. In a nutshell, any evaluation of FL must weigh its benefits against its costs. Focusing exclusively on the growth effects of liberalization during good times would miss the link between FL and crises. Focusing only on volatility and crises could lead to an excessive cautiousness about the risks of FL. The case for FL requires that its growth and welfare benefits outweigh the costs associated with more frequent financial crises.

Measuring Financial Liberalization

There are three classes of FL indices. First, there are *de jure* indices based on official dates of policy reforms. An example is the index based on the IMFAnnual Report on *Exchange Arrangements* and *Exchange Restrictions* (Grilli and Milesi-Ferretti 1995). This class of indices permits a comparison of the periods before and after liberalization. A drawback, however, is that legislated changes take time to translate into liberalization on the ground. Liberalization may even fail to materialize altogether when well-functioning domestic financial markets are absent. Bekaert et al. (2005) overcome this problem by constructing a de jure indicator of equity market liberalization that records the date after which foreign investors are able to invest in domestic securities. A second class of indices uses *de facto* measures of financial openness, like the capital flows–GDP ratio used by Edison et al. (2004). The drawback is that these measures are contaminated by cyclical fluctuations and thus are imprecise indicators for dating FL. Lastly, *de facto* indices identify structural breaks in the trend of capital inflows (Tornell et al. 2003). These indices combine the advantages of the two previous classes as they provide more precise FL dates based on actual, rather than merely legislated, policy reforms.

Financial Liberalization and Growth

BHL (2005) find that equity market liberalization leads to an increase of one percentage point in average real per-capita GDP growth. Rancière et al. (2005) find that capital account liberalization leads to a similar gain in growth. To illustrate the link between FL and growth we add liberalization dummies to a standard growth regression:

$$\Delta y_{it} = \lambda y_{i,\,\mathrm{ini}} + \gamma X_{it} + \varphi_1 TL_{it} + \varphi_2 FL_{it} + \varepsilon_{jt}, \quad (1)$$

where Δy_{it} is the average growth rate of GDP per capita; $y_{i,\mathrm{ini}}$ is the initial level of GDP per capita; X_{it} is a vector of control variables that includes initial human capital, the average population growth rate, and life expectancy; and TL_{it} and FL_{it} are the trade and financial liberalization dummies of TWM (2003), respectively. For each country and each variable, we construct 10-year averages starting with the period 1980–1989 and rolling forward to the period 1990–1999. Thus each country has up to ten data points in the time-series dimension. The liberalization dummies take values in the interval [0,1], depending on the proportion of liberalized years in a given window. We estimate the panel regressions using generalized least squares.

F

Financial Liberalization, Table 1 Regressions explaining growth in GDP per capita, 1980–1999[a]

Independent variable[a]	1–1	1–2	1–3	1–4	1–5[b]	1–6	1–7[b]
Mean of real credit				0.154[c]	0.170[c]	0.110[c]	0.093[c]
growth rate				(0.009)	(0.012)	(0.009)	(0.007)
Standard deviation of				−0.030[c]	−0.029[c]	−0.019[c]	−0.014[c]
real credit growth rate				(0.003)	(0.007)	(0.004)	(0.003)
Negative skewness of real				0.266[c]	0.174[c]	0.135[c]	−0.095[c]
credit growth rate				(0.021)	(0.069)	(0.031)	(0.053)
Financial liberalization	1.530[c]		1.443[c]		1.811[c]	1.894[c]	
	(0.191)		(0.221)		(0.163)	(0.122)	
Trade liberalization		0.793[c]	0.776[c]			0.895[c]	0.838[c]
		(0.152)	(0.196)			(0.198)	(0.155)
Summary statistics:							
Adjusted R^2	0.848	0.897	0.807	0.629	0.667	0.731	0.752
No. of observations	409	430	408	424	269	408	253

Notes: [a]The estimated equations are Eqs. (1) and (2) in the text; the dependent variable is the average annual growth rate of real GDP per capita. Control variables include initial per capita income, secondary schooling, population growth, and life expectancy. Standard errors are reported in parentheses and are adjusted for heteroskedasticity according to Newey and West (1987)

[b]This regression includes the group of middle-income countries only

[c]Significance at the 5% level. The equation is estimated in an overlapping panel regression by GLS with data as ten-year averages starting with 1980–1989 and rolling forward to 1990–1999

Source: Authors' regressions

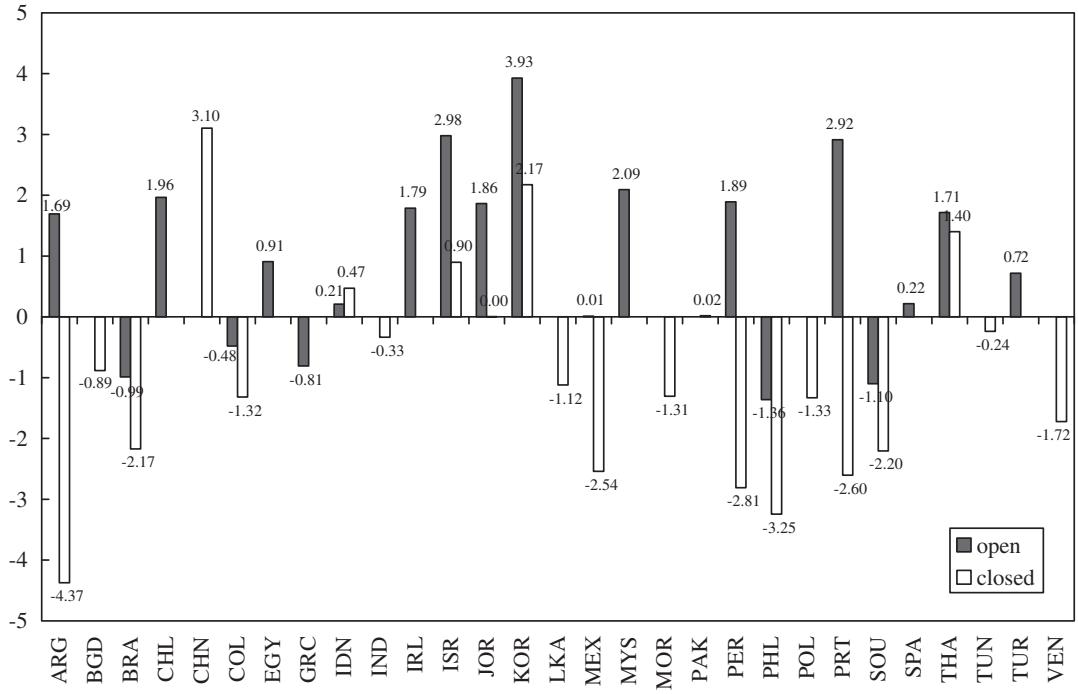
The FL dummy enters significantly at the 5% level in all regressions. Regression 1–1 of Table 1 shows that following FL growth in real GDP per-capita increases by 1.5 percentage points per year, after controlling for the standard variables. Trade liberalization increases growth by 0.8% per year (column 1–2). When both liberalization dummies are included (column 1–3), both enter significantly. This suggests that trade and financial liberalization have independent effects and jointly contribute to higher long-run growth.

Figure 2 illustrates the link between FL and growth for individual MICs. For each country, we plot growth residuals before and after FL. Growth residuals are obtained by regressing real per capita growth on initial income per capita and population growth. Figure 2 shows clearly that for almost all countries growth has been higher in the financially liberalized period.

Several studies find mixed evidence on the link between financial openness and growth. This can be attributed either to the indicators of openness used or to the sample considered. First, some studies include low-income countries that do not have functioning financial markets. In these countries we do not expect the financial deepening mechanism to work. One might also expect the growth effect of FL to be smaller in high-income than in middle-income countries as the latter face more severe borrowing constraints. Hence, sample heterogeneity can create a bias against finding a linear growth effect of FL. Klein (2005) finds that FL contributes to growth among MICs but not among poor or rich countries. Second, some studies test the effect of changes in the capital flows–GDP ratio on growth. However, because this index does not identify a specific liberalization date, it is not appropriate for comparing the behaviour of macroeconomic variables before and after liberalization. Furthermore, these measures tend to exhibit year-to-year fluctuations that do not reflect actual changes in the degree financial openness.

Financial Liberalization and Crises

FL is typically followed by boom–bust cycles. During the boom, bank credit expands very rapidly and excessive credit risk is undertaken. As a result, the economy becomes financially fragile and prone to crisis. Although the likelihood that



Financial Liberalization, Fig. 2 Liberalization and annual percent growth (*Note*: The country episodes are constructed using windows of different length for each country. Country episodes that are shorter than 5 years are excluded. Averaging over these periods, we estimate a simple growth regression by OLS in which real per capita growth is the dependent variable and which only includes the respective initial income and population growth. The figure plots the residuals from this regression, from 1980 to 1999. *Sources:* Population growth for Portugal: International Financial Statistics (IMF). All other series: World Bank Development Indicators)

a lending boom will crash in a given year is low, many booms do eventually end in a crisis. During such a crisis, new credit falls abruptly and recuperates only gradually.

The incidence of crises can be measured by analysing countries' financial histories and by codifying the occurrence of banking crises, currency crises, and sudden stops in capital inflows. Kaminsky and Reinhart (1999) use such a crisis index in a probit model to test whether banking and currency crises are crises are more likely to occur after FL.

RTW (2005) use a more parsimonious indicator of financial fragility: the negative skewness of credit growth. Negative skewness is a de facto indicator that captures the existence of infrequent, sharp and abrupt falls in credit growth. Since credit growth is relatively smooth during boom periods, and crises happen only occasionally, in financially fragile countries the distribution of credit growth rates is characterized by negative outliers in a long enough sample. These outliers correspond to the abrupt falls in credit growth that occur during the crisis or 'bust' stage of the boom–bust cycle. The advantages of this skewness measure, relative to other more complex indicators of crises, are that it is objective and comparable across countries.

In the literature variance is the typical measure of volatility. Variance, however, is not a good instrument to identify growth-enhancing credit risk because high variance reflects not only the presence of boom–bust cycles but also the presence of high-frequency shocks.

Table 2 partitions country-years into two groups: liberalized and non-liberalized. The table shows that, across MICs, the financial deepening induced by FL has not been a smooth process but has been characterized by booms and occasional busts. We can see that FL leads to an increase in

Financial Liberalization, Table 2 Moments of credit growth before and after financial liberalization

Moment	Liberalized country-years	Non-liberalized country-years
MICs		
Mean	0.078	0.038
Standard deviation	0.151	0.170
Skewness	−1.086	0.165
HICs		
Mean	0.025	. . .
Standard deviation	0.045	. . .
Skewness	0.497	. . .

Note: The sample is partitioned into two country-year groups: liberalized and non-liberalized. Before the standard deviation and skewness are calculated, the means are removed from the series and data errors for Belgium, New Zealand and the United Kingdom are corrected for. The total sample ranges from 1980 to 1999
Source: Authors' calculations

the mean of credit growth of four percentage points (from 3.8 to 7.8%) and a fall in the skewness of credit growth from near zero to −1.09, and has only a negligible effect on the variance of credit growth. Notice that, across high-income countries, credit growth exhibits near-zero skewness, and both the mean and the variance are smaller than across MICs. This difference reflects the absence of severe credit market imperfections in high-income countries.

Growth and Crises

To close the circle we show that countries with a greater incidence of crises countries have grown faster than those with smooth credit paths. We do so by adding three moments of real credit growth to growth regression (1)

$$\Delta y_{it} = \lambda y_{i,\mathrm{ini}} + \gamma X_{it} + \beta_1 \mu_{\Delta B,it} + \beta_2 \sigma_{\Delta B,it} + \beta_3 S\Delta_{B,it} + \varphi_1 TL_{it} + \varphi_2 FL_{it} + \varepsilon_{jt}, \quad (2)$$

where Δy_{it}, $y_{i,\mathrm{ini}}$, X_{it}, TL_{it}, and FL_{it} are defined as in Eq. (1), and $\mu_{\Delta B,it}$, $\sigma_{\Delta B,it}$, and $S_{\Delta B,it}$ are the mean, standard deviation, and skewness of the real credit growth rate, respectively. We estimate Eq. (2) using the same type of overlapping panel data regression as for Eq. (1). Columns 1–4 through 1–7 of Table 1 report the estimation results. Consistent with the literature, we find that, after controlling for the standard variables, the mean growth rate of credit has a positive effect on long-run GDP growth, and the variance of credit growth has a negative effect. Both variables enter significantly at the 5% level in all regressions.

The first key point is that the financial deepening that accompanies rapid GDP growth is not smooth but, rather, takes place via booms and busts. Columns 1–4 and 1–5 show that negative skewness – a bumpier growth path – is on average associated with faster GDP growth across countries with functioning financial markets. This estimate is significant at the 5% level.

To interpret the estimate of 0.27 for skewness, consider India, which has nearzero skewness, and Thailand, which has a skewness of about minus 2. A point estimate of 0.27 implies that an increase in the bumpiness index of 2 (from zero to minus 2) increases the average long-run GDP growth rate by 0.54 of a percentage point a year. Is this estimate economically meaningful? To address this question, note that, after controlling for the standard variables, Thailand grows about two percentage points faster per year than India. Thus, about a quarter of this growth differential can be attributed to credit risk taking, as measured by the skewness of credit growth.

The second key point is that the association between skewness and growth does not imply that crises are good for growth. Crises are costly. They are the price that has to be paid in order to attain faster growth in the presence of credit market imperfections. To see this, consider column 1–6. When the FL dummy is included, bumpiness enters with a negative sign (and is significant at the 5% level). In the MIC set, given that there is FL, the lower the incidence of crises the better. We can see the same pattern when we include high-income countries in column 1–7.

Clearly, liberalization without fragility is best, but the data suggest that this combination is not available to MICs. Instead, the existence of contract enforceability problems implies that liberalization leads to higher growth because it eases

financial constraints but, as a by-product, also induces financial fragility. However, because crises occur relatively rarely, FL has a positive net effect on long-run growth.

A Unified Approach

An alternative approach to understand the contrasting effects of FL is to combine the linear growth regression with a crisis probit model. In this way one can decompose the net effect of FL into a direct pro-growth effect and an indirect anti-growth effect, via a higher propensity to crises. Using this approach, RTW (2006) find that the direct effect of FL on growth is 1.2 percentage points and the indirect effect is minus 0.25 percentage points. In order to understand this result, one should keep in mind that even in financially liberalized countries crises are rare events. Therefore, even if crises have large output consequences, their estimated growth effect remains modest. In contrast, since FL is likely to improve access to external finance, it has a firstorder impact on growth.

Conceptual Framework

To analyze FL and the subsequent boom–bust cycles, consider an economy with two sectors: non-tradables (N) and tradables (T). Alternatively, one can think of 'neweconomy' and 'traditional' sectors, respectively. The key is that each sector uses as input the other sector's output.

This economy is subject to severe contract enforceability problems that generate financing constraints. While T-firms can overcome such constraints and finance themselves in bond and equity markets, most N-firms are financially constrained and bank-dependent. Since N-goods serve as intermediate inputs for both sectors, the N-sector constrains the long-run growth of the T-sector and that of GDP: there is a bottleneck.

In such an economy, FL increases GDP growth by increasing the investment of financially constrained firms. However, the easing of financial constraints is associated with the undertaking of insolvency risk because FL not only lifts restrictions that preclude risk taking but also is associated with explicit and implicit systemic bail-out guarantees that cover creditors against systemic crises.

It is a stylized fact that, if a critical mass of borrowers is on the brink of bankruptcy, authorities will implement policies to ensure that creditors get repaid (at least in part) and thus avoid an economic meltdown. These bail-out policies may come in the form of an easing of monetary policy in response to a financial crash, the defence of an exchange rate peg in the presence of liabilities denominated in foreign currency, or the recapitalization of the financial sector.

Because domestic banks have been the prime beneficiaries of these guarantees, investors use domestic banks to channel resources to firms that cannot pledge international collateral. Thus liberalization results in biased capital inflows. T-firms and large N-firms are the recipients of foreign direct investment (FDI) and portfolio flows, whereas most of the inflows to the N-sector are intermediated through domestic banks, which enjoy bail-out guarantees. Insolvency risk often takes the form of maturity mismatch or risky debt denomination (currency mismatch).

Taking on insolvency risk reduces expected debt repayments because authorities will cover part of the debt obligation in the event of a systemic crisis. Thus the guarantee allows financially constrained firms to borrow more than they could otherwise. This increase in borrowing and investment is accompanied by an increase in insolvency risk. When many firms take on insolvency risk, aggregate financial fragility arises together with increased N-sector investment and growth. Faster Nsector growth then helps the T-sector grow faster because N-sector goods are used in T-sector production. Therefore, the T-sector will enjoy more abundant and cheaper inputs than otherwise. As a result, as long as a crisis does not occur, growth in a liberalized economy is faster than in a non-liberalized one.

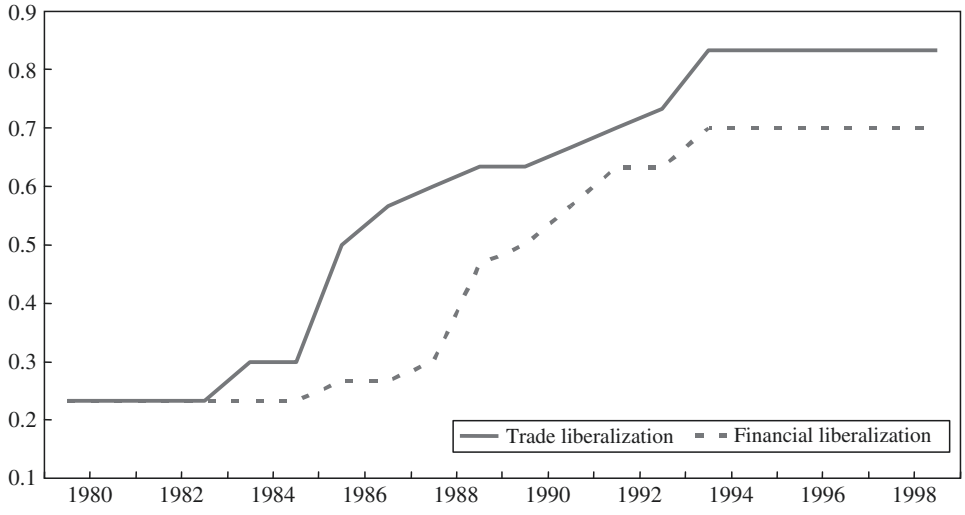
Of course, financial fragility implies that a self-fulfilling crisis may occur. And during crises GDP growth falls. Crises must be rare, however, in order to occur in equilibrium – otherwise agents

would not find it profitable to take on credit risk in the first place. Thus, average long-run growth is greater along a risky path than along a safe one even if there are large crisis costs. This is why FL leads both to higher long-run growth and to a greater incidence of crises. Schneider and Tornell (2004) and RTW (2003) formalize the intuitive argument we described using a general equilibrium model with rational agents.

This discussion of the mechanism through which FL affects the growth of MICs also explains why FL does little to improve the growth of LICs. LICs often do not have functioning financial markets and thus lack the infrastructure that allows the financial system to direct international funds to profitable firms. MICs, by contrast, have enough financial infrastructure to allocate funds reasonably well, even though contract enforceability problems prevent them from doing so as efficiently as high-income countries (HICs). Because of the imperfections in their financial systems, the price of fast growth in MICs is financial fragility. The contrasting experiences of Thailand and India during the period 1980–2002 illustrate this trade-off clearly. As we discussed earlier, Thailand experienced booms and busts while India did not. While Thailand experienced spectacular growth, India's growth was dismal. Recently, India has opened its economy to both trade and finance. Not surprisingly, India is currently experiencing a lending boom. It will be interesting to analyse the evolution of the Indian economy around 2015.

Economic Policy

Several observers have suggested that partial liberalization is the optimal policy to reap the growth benefits of openness without having to suffer from volatility and crises. They suggest the implementation of trade liberalization but not of FL, or the restriction of capital flows to FDI, the least volatile form of capital flows. These recommendations seem impractical. First, an open trade regime is usually sustained by an open financial regime because exporters and importers need access to international financial markets. Since capital is fungible, it is difficult to insulate the financial flows associated with trade transactions. The data indicates that trade liberalization has typically been followed by FL. As Fig. 3 shows, by 1999 72% of countries that had liberalized trade had also liberalized financial flows, bringing the share of MICs that were financially liberalized to 69%, from 25% in 1980.



Financial Liberalization, Fig. 3 Share of MICs that liberalized trade and financial flows, 1980–1999 (*Note*: The figure shows the share of countries that have liberalized relative to the total number of MICs in our sample. *Source*: Tornell et al. (2003))

Second, FDI does not obviate the need for risky international bank flows. FDI goes mostly to financial institutions and large firms, which are mostly T-firms. Thus, bank flows are practically the only source of external finance for most N-firms (Tornell and Westermann 2005). Curtailing such risky flows would reduce N-sector investment and generate bottlenecks that would limit long-run growth. Bank flows are hardly to be recommended, but for most firms it might be that or nothing. Clearly, allowing risky capital flows does not mean that anything goes. Appropriate prudential regulation must also be in place.

In an environment with asymmetric financial opportunities authorities may be tempted to make direct investment subsidies to constrained sectors. The historical evidence indicates that such centrally planned policies typically fail. We now know that either authorities do not possess the appropriate information or crony capitalism and rampant corruption take over. A second-best policy is to liberalize financial markets and allow banks to be the means through which resources are channelled to financially constrained firms. Here, it is important to make a distinction between 'systemic' and 'unconditional' bail-out guarantees. The former are granted only if a critical mass of agents default. The latter are granted on an idiosyncratic basis whenever there is an individual default. We have argued that, if authorities can commit to grant only systemic guarantees, and if prudential regulation works efficiently, then FL will induce higher long-run growth in a credit-constrained economy. In contrast, if guarantees are granted on an unconditional basis or there is a lax regulatory framework, the monitoring and disciplinary role of banks in the lending process will be negated. In this case, FL will simply lead to overinvestment and corruption.

One should not conclude that in order to enjoy the growth and welfare benefits of FL countries have to be exposed for ever to the risk of crises. The amelioration of contract enforceability problems, through a better legal system and other institutional reforms, is a fundamental source of higher growth and lower volatility in the long-run. However, it often takes time for these reforms to be achieved. In the meantime, countries with functioning financial markets can be made better off by liberalizing and experiencing a rapid but risky growth path, rather than remaining closed and trapped in a safe but slow growth path.

See Also

- ▶ Banking Crises
- ▶ Currency Crises
- ▶ Foreign Direct Investment
- ▶ International Capital Flows

F

Bibliography

Bekaert, G., C. Harvey, and C. Lundblad. 2005. Does financial liberalization spur growth? *Journal of Financial Economics* 77: 3–55.

Edison, H., M. Klein, L. Ricci, and T. Sløk. 2004. Capital account liberalization and economic performance: Survey and synthesis. *IMF Staff Papers* 51: 111–155.

Grilli, V., and G. Milesi-Ferretti. 1995. *Economic effects and structural determinants of capital controls*, Working Papers No. 95/31. Washington, DC: IMF.

Kaminsky, G., and C. Reinhart. 1999. The twin crises: The causes of banking and balance-of-payments problems. *American Economic Review* 89: 473–500.

Klein, M. 2005. *Capital account liberalization, institutional quality and economic growth: Theory and evidence*, Working Paper No. 11112. Cambridge, MA: NBER.

Newey, W., and K. West. 1987. A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55: 703–708.

Rancière, R., A. Tornell, and F. Westermann. 2003. *Crises and growth: A reevaluation*, Working Paper No. 10073. Cambridge, MA: NBER.

Rancière, R., A. Tornell, and F. Westermann. 2005. *Systemic crises and growth*, Working Paper No. 11076. Cambridge, MA: NBER.

Rancière, R., A. Tornell, and F. Westermann. 2006. Decomposing the effects of financial liberalization: Crises vs. growth. *Journal of Banking and Finance* 30: 3331–3348.

Schneider, M., and A. Tornell. 2004. Balance sheet effects, bailout guarantees and financial crises. *Review of Economic Studies* 71: 883–913.

Tornell, A., and F. Westermann. 2005. *Boom-bust cycles and financial liberalization*. Cambridge, MA: MIT Press.

Tornell, A., F. Westermann, and L. Martinez. 2003. Liberalization, growth and financial crisis: lessons from Mexico and the developing world. *Brookings Papers on Economic Activity* 2: 1–112.

Financial Market Anomalies

Donald B. Keim

Abstract

Financial market anomalies are cross-sectional and time series patterns in security returns that are not predicted by a central paradigm or theory. The focus here is on equity market anomalies including the size effect, value effect, serial correlation in returns and calendar-related patterns in returns related to month of the year and day of the week. Many of these patterns have persisted for decades, suggesting they are not evidence of market inefficiencies. Although transactions costs might preclude trading that would eliminate such patterns, it is possible that our benchmark models might be less than complete descriptions of equilibrium price formation.

Keywords

After-tax asset pricing models; Bid-ask spread; Capital asset pricing model; Capital gains tax; Cross section of stock returns; Dividend yield effect; Equity premium puzzle; Financial market anomalies; Informational efficiency; Kuhn, T.; Liquidity effect; Measurement error; Momentum effect; Risk premia; Size effect; Taxation of income; Time series patterns in security returns; Value effect; Weekend effect

JEL Classifications

G1

Financial market anomalies are cross-sectional and time series patterns in security returns that are not predicted by a central paradigm or theory. This sense of the term 'anomaly' can be traced to Kuhn (1970). Documentation of anomalies often presages a transitional phase towards a new paradigm.

Discoveries of financial market anomalies typically arise from empirical tests that rely on a joint null hypothesis – to wit, security markets are informationally efficient *and* returns behave according to a pre-specified equilibrium model (for example, the capital asset pricing model, CAPM). If the joint hypothesis is rejected, we cannot attribute the rejection to either branch of the hypothesis. Thus, even though anomalies are often interpreted as evidence of market inefficiency, such a conclusion is inappropriate because the rejection may be due to an incorrect equilibrium model. Some have argued that, once identified by researchers, the magnitude of financial anomalies will tend to dissipate as investors seek to profitably exploit the return patterns or because their discovery was simply a sample-specific artifact. Although this has happened for some of the findings discussed below (such as the weekend effect), most of the anomalies discussed continue to persist. The fact that so many of these patterns have persisted for decades suggests that they are not evidence of market inefficiencies. Rather, our benchmark models might be less than complete descriptions of equilibrium price formation.

The number of documented anomalies is large and continues to grow. The focus here is on equity market anomalies, and on the subset whose existence has proven most robust with respect to both time and the number of stock markets in which they have been observed. We broadly classify the findings as being cross-sectional or time series in nature.

Cross-Sectional Return Patterns

Given certain simplifying assumptions, the CAPM states that the return on a security is linearly related to the security's non-diversifiable risk (or beta) measured relative to the market portfolio of all marketable securities. If the model is correct and security markets are efficient, security returns will on *average* conform to this linear relation.

Empirical tests of the CAPM first became possible with the creation of computerized databases of stock prices in the United States in the 1960s. To implement the tests, researchers often estimate cross-sectional regressions of the form

$$R_i = a_0 + a_1\beta_i + \sum a_j c_{ij} + e_i \qquad (1)$$

where β_i is the security's beta which measures its covariance with the return on the market and c_{ij} represents security-specific characteristic j (size, earnings yield, and so on) for security i. The CAPM predicts that the a_j, for $j > 1$, are zero. Early tests supported the CAPM (for example, significant positive values for a_1, insignificant values for a_j, for $j > 1$). The explanatory power of beta came into question in the late 1970s when researchers identified security characteristics such as the earnings-to-price ratio and market capitalization of common equity with more explanatory power than beta.

This section presents a sample of the more important contributions in this area that collectively stand as a challenge for alternative asset pricing models.

The Value Effect

The value effect refers to the positive relation between security returns and the ratio of accounting-based measures of cash flow or value to the market price of the security. Examples of the accounting-based measures are earnings per share and book value of common equity per share. Investment strategies based on the value effect have a long tradition in finance and can be traced at least to Graham and Dodd (1940). Ball (1978) argues that variables like the earnings-to-price ratio (E/P) are proxies for expected returns. Thus, if the CAPM is an incomplete specification of priced risk, it is reasonable to expect that E/P might explain the portion of expected return that is compensation for risk variables omitted from the tests.

Basu (1977) was the first to test the notion that value-related variables might explain violations of the CAPM. He found a significant positive relation between E/ P ratios and average returns for US stocks that could not be explained by the CAPM. Reinganum (1981) confirmed and extended Basu's findings. Rosenberg et al. (1985), De Bondt and Thaler (1987) and many others have documented a significant positive relation between returns and the book-to-price ratio (B/P). Researchers have also identified a significant relation between security returns and value ratios that use cash flow (earnings plus accounting depreciation expense) in place of earnings in the numerator of the ratio. The value effect in its many forms has been reproduced by numerous researchers for many different sample periods and for most major securities markets around the world (see Hawawini and Keim 2000, for a review).

Dividend yield, the ratio of cash dividend to price, has also been shown to have cross-sectional return predictability. Although similar in construction to the value ratios, the explanatory power of dividend yields is most often attributed to the differential taxation of capital gains and ordinary income as described in the after-tax asset pricing models developed by Brennan (1970) and Litzenberger and Ramaswamy (1979). Although a positive relation between stock returns and dividend yields has been documented in many studies, interpretation of the results as support for an after-tax pricing model has been controversial. Evidence on the dividend yield effect has been provided by Litzenberger and Ramaswamy (1979), Miller and Scholes (1982) and many others.

The Size Effect

The size effect refers to the negative relation between security returns and the market value of the common equity of a firm. Banz (1981) was the first to document this phenomenon for US stocks (see also Reinganum 1981). In the context of Eq. (1), Banz found that the coefficient on size has more explanatory power than the coefficient on beta in describing the cross section of returns. Indeed, Banz finds little explanatory power for market betas. Like the value effect, the size effect has been reproduced for numerous sample periods and for most major securities markets around the world (Hawawini and Keim 2000).

Interpretation of the Value and Size Effects

The separately identified value and size effects are not independent phenomena because the security characteristics all share a common variable – price per share of the firm's common stock. Indeed, researchers have shown a high rank correlation between size and price and between the value ratios and price, and others have documented a

significant cross-sectional relation between price per share and average returns. To sort out the relative importance of the different variables, Fama and French (1992) (FF) estimate Eq. (1) with multiple value and size variables included as explanatory variables (see also Jaffe et al. 1989). FF find that B/ P and Size provide the greatest explanatory power in describing the cross section of returns, and suggest that B/P and Size are proxies for the influence of two additional risk factors omitted from the CAPM. In this context, the value and size variables can be viewed as capturing sensitivities to the omitted factors, and the coefficients multiplying the value and size variables (a_j in Eq. (1)) are estimates of the risk premia required to compensate for that exposure. (A valid question is whether a characteristic like B/P proxies for an underlying (but unknown) risk factor which is the determinant of expected returns or whether the characteristic itself is the determinant of expected returns. Daniel and Titman 1997, address this issue and conclude that security characteristics appear to be more important than the covariance of security returns with a factor related to the characteristic.) Predicated on this interpretation, Fama and French (1993) propose a three-factor model to describe the time series behaviour of security returns:

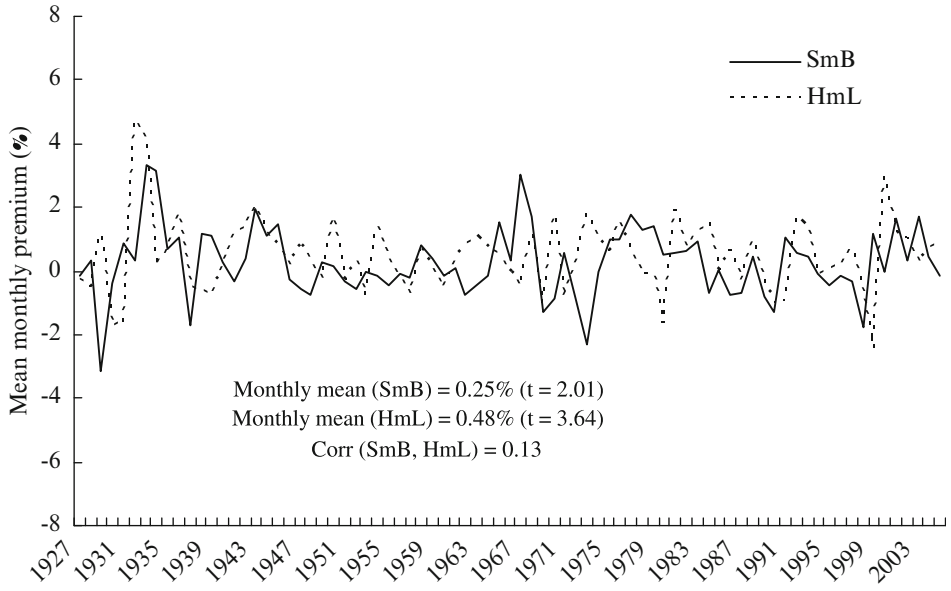
$$R_t - r_{f,t} = \beta_0 + \beta_1\left(r_{m,t} - r_{f,t}\right) + \beta_2 SmB_t + \beta_3 HmL_t + \varepsilon_t \quad (2)$$

where R_t is the return on the asset in month t, $r_{f,t}$ is the monthly treasury bill rate, $r_{m,t}$ is the return on a value-weighted market portfolio, SmB_t is a monthly size premium (Small stock return minus Large stock return), HmL_t is a monthly value premium (High B/P return minus Low B/P return), and ε_t is the error term. As constructed, *SmB* and *HmL* are zero net investment portfolios. If these three factors span all sources of common systematic co-movement in security returns, β_0 ('alpha') will on average equal zero. The model has received much empirical confirmation and appears to explain numerous previously reported incidences of anomalous cross-sectional return patterns (that is, such effects have $\beta_0 = 0$ in Eq. (2)).

As mentioned above, the mean values of the three factors in model (2) can be interpreted as the premium or compensation earned by an investment position for unit exposure to each separate factor. The relative magnitudes of these factor premia are of economic interest. The market risk premium quantifies the return, in excess of a default-risk-free return, provided for investing in a broadly diversified portfolio as represented by the value-weighted market portfolio. Over the period 1927–2005 the average equity market risk premium in (2) is 0.64 per cent per month. Utility-based asset pricing models have difficulty explaining an equity premium of this magnitude–either because the returns on default-risk-free bonds are too low, or the returns on equities are too high. This has been called the equity premium puzzle (Mehra and Prescott 1985) and has generated an extensive literature trying to reconcile the theory and empirical evidence.

The mean risk premia associated with the size effect (*SmB*) and the value effect (*HmL*) should be zero if the CAPM is correct. Consistent with the research described above, *SmB* and *HmL* are both positive. For the period January 1927–December 2005 the monthly mean (*t*-value) is 0.25 per cent (2.01) for *SmB* and 0.48 per cent (3.64) for *HmL*, and the correlation between the two premia is 0.13. Figure 1 plots the time series of the intra-year monthly means of the two premia. The figure shows that (*a*) both premia display substantial variability over time and (*b*) the two series display a considerable common co-movement despite the low estimated correlation.

On the first point, there are extended periods when the signs of the risk premia are reversed. This is particularly evident for the size effect – for extended periods in the 1950s and the 1980s large firms outperformed small firms, in contrast to other periods (1930s, 1940s, 1970s, and post-2000) when small stocks outperformed large stocks. Because the estimated magnitudes of the effects are sensitive to the period in which they are measured, it is important to distinguish between unconditional and conditional expected values for the effects. Further, it is relevant to ask whether the 79-year sample we have for the US market (longer than in other developed equity



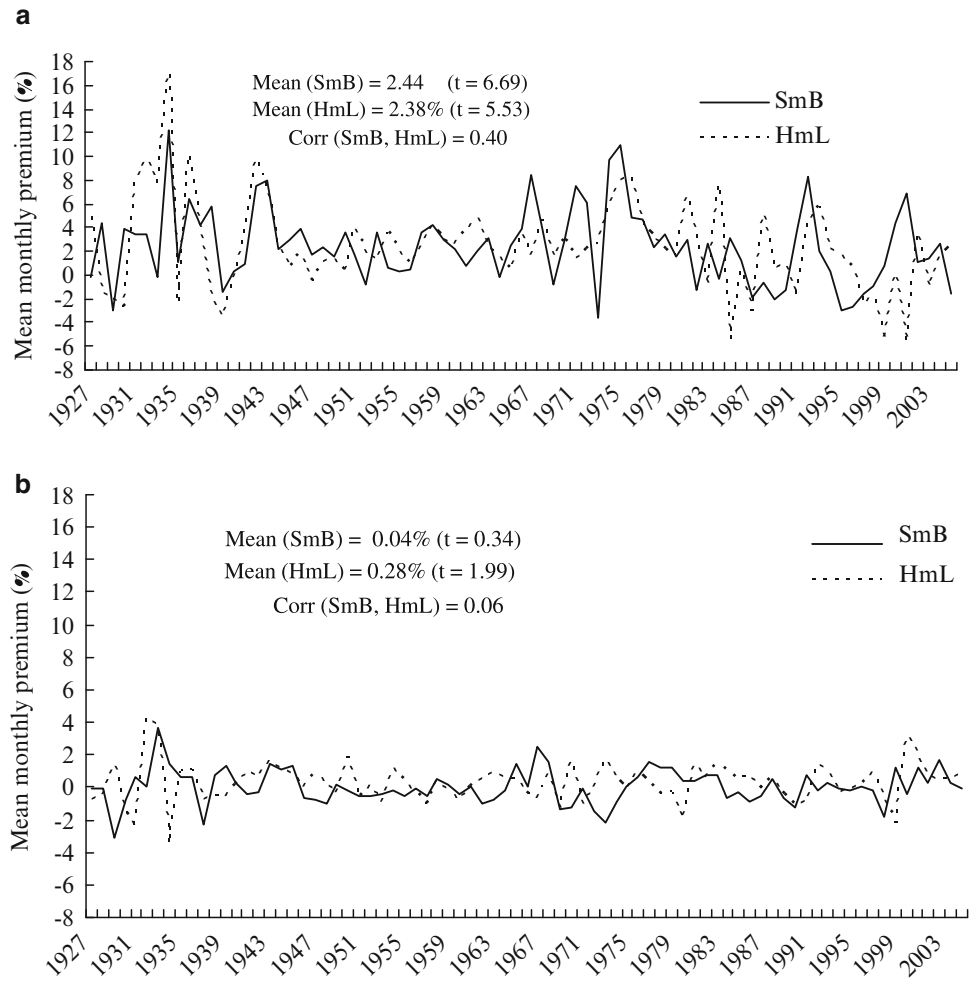
Financial Market Anomalies, Fig. 1 The value and size premia, 1927–2005

markets) is long enough to capture the 'long-run' magnitudes of such volatile effects. (The same caveat has been raised regarding the magnitude of the equity premium.)

On the second point, the visual appearance of common co-movement between the series suggests the two effects are not entirely independent. This possibility is confirmed when the time series plots of *SmB* and *HmL* are decomposed into separate plots for January and February-to-December observations (Fig. 2a, b). Much research has shown that the size and value effects are most pronounced in the month of January. This research is discussed in more detail in the next section. For now, we limit discussion to the difference in the behaviour of *SmB* and *HmL* between January and February–December. First, the mean values for both premia are an order of magnitude larger in January than in February–December. Second, the correlation of 0.40 in January versus 0.06 for February– December demonstrates that the commonality between the two series in Fig. 1 arises mostly from their common behaviour in January.

What explains the value and size effects? That both premia reflect some common element which manifests itself only in January is hard to reconcile with a risk compensation story. (Non-risk-based explanations of the January effect are discussed in the section on seasonal patterns in stock returns below.) Much recent research, nevertheless, has characterized the value premium as compensation for financial distress risk. Theoretical models have been developed in which such risk plays a central role, and value (high B/P) stocks accordingly earn higher equilibrium returns than growth (low B/P) stocks. Others have argued that the size effect is actually a liquidity effect in which small-cap stocks are less liquid than large-cap stocks and therefore provide correspondingly higher returns to offset the higher transactions costs (see, for example, Brennan et al. 1998). Still others have suggested that the size and B/P results may be due to survivor biases in the databases used by researchers (see, for example, Kothari et al. 1995).

One final hypotheses concerns measurement error in the estimated market betas used in the tests. Firms whose stocks have recently declined in price (for example, many high B/P and small-cap stocks), in the absence of a concomitant decline in the value of the debt, have become more leveraged and, other things equal, more risky in a beta sense. Traditional estimation



Financial Market Anomalies, Fig. 2 (**a**) The value and size premia – January only. (**b**) The value and size premia – Feb to Dec

methods produce 'stale' betas that underestimate 'true' beta risk for such firms. Thus, B/P and size may be viewed as better instruments for 'true' market beta risk than traditional estimates of beta, and the value and size effects are simply capturing the measurement error in the traditional beta estimates.

The Prior Return or Momentum Effect

Prior stock returns have been shown to have explanatory power in the cross section of common stock returns. Stocks with prices on an upward (downward) trajectory over a prior period of 3 to 12 months have a higher than expected probability of continuing on that upward (downward) trajectory over the subsequent 3 to 12 months. This temporal pattern in prices is referred to as momentum. Jegadeesh and Titman (1993) show that a strategy that simultaneously buys past winners and sells past losers generates significant abnormal returns over holding periods of 3 to 12 months. The abnormal profits generated by such offsetting long and short positions appear to be independent of market, size or value factors, and have persisted in the data for many years. To this end, Carhart (1997) estimates an extension of Eq. (2) that includes a momentum factor (in addition to market, size and value factors)

defined in the spirit of Jegadeesh and Titman as the difference in returns between a portfolio of 'winners' and a portfolio of 'losers'. The coefficient on the momentum factor is positive and statistically significant, and cannot be explained by the other three factors. Finding a rational risk-related explanation for the momentum effect has proven difficult. A number of researchers have posited behavioural (psychology-based) explanations of momentum that rely on irrational market participants who underreact to news, but these models are hard to reconcile with psychology-based models of overreaction posited to explain the value premium (for example, Lakonishok et al. 1994).

Time Series Return Predictability

Consider a model of stock prices in which expected stock returns are constant through time (see Fama 1976, for discussion of this model and related tests of the behaviour of stock prices). Much recent evidence suggests that expected returns are not constant, but contain a time-varying component that is predicted by past returns, *ex ante* observable variables and calendar turning points. The following subsections discuss this evidence.

Predicting Returns with Past Returns I: Individual Security Autocorrelations

Much research finds that autocorrelations of higher-frequency (daily, weekly) individual stock returns are negative and that the autocorrelations are inversely related to the market capitalization of the stock. The exception is that the largest market cap stocks have positive autocorrelations for daily returns. The inverse relation between individual return autocorrelations and market capitalization is due to the influence of a bid-ask bounce in high frequency stock prices that may induce 'artificial' serial dependencies into returns. Niederhoffer and Osborne (1966) find that successive trades tend to occur alternately at the bid and then the ask price, resulting in negative serial correlation in returns. This negative serial dependency is more pronounced for smaller stocks that have lower prices and, consequently, for which the bid-ask spread represents a larger percentage of price. Because of the high variance of individual stock returns, researchers find that past returns explain a trivial percentage of total return variability at high frequencies (typically less than one per cent). And the predictability at high frequencies is economically insignificant: profits from trading strategies attempting to exploit the predictability in individual stocks are indistinguishable from zero.

Predicting Returns with Past Returns II: Aggregate Return Autocorrelations

Because of variance reduction obtained from diversification, aggregated or portfolio returns provide more powerful tests of return predictability using past returns. However, this increased power may be offset by upward-biased autocorrelations caused by the infrequent trading of securities in the portfolios (Fisher 1966). This bias is more serious for portfolios of smaller-cap stocks that contain less frequently traded stocks. In the United States and other global equity markets positive autocorrelations for high-frequency portfolio returns range from 0.4 for small-cap stocks to 0.1 for large-cap stocks. Research has shown, however, that positive portfolio autocorrelations are not due to infrequent trading of the securities in the portfolio. Indeed, many researchers have reported statistically significant positive portfolio autocorrelations for return frequencies up to one month in the United States, an interval over which virtually all securities will have traded. There is no evidence, however, of profitable trading opportunities based on daily, weekly or monthly aggregate return autocorrelations. (Lo and MacKinlay 1990, reconcile the paradox of positive portfolio autocorrelations and negative individual stock autocorrelations: because the autocorrelation of portfolio returns is the sum of individual security autocovariances and cross-autocovariances, if the cross-autocovariances are sufficiently larger than the autocovariances – empirically, they are – then the cross-autocovariances will overshadow the contribution of the autocovariances.)

Significant predictability – both economically and statistically – has been identified in longer-horizon stock returns. As mentioned in the

previous section, Jegadeesh and Titman (1993) identify profitable trading strategies based on past price momentum over 3-to 12-month intervals. De Bondt and Thaler (1985) find that New York Stock Exchange stocks identified as the biggest losers (winners) over a period of three to five years earn, on average, the highest (lowest) market-adjusted returns over a subsequent holding period of the same length of time, a phenomenon that does not seem to disappear when returns are adjusted for size and beta risk. This predictable reversal pattern is often attributed to market 'overreaction' in which stock prices diverge from fundamental values because of (irrational) waves of optimism or pessimism before returning eventually to fundamental values. Evidence of this longer-horizon return predictability has been reported in most equity markets around the world. But the significance of negative autocorrelation for long horizon returns is subject to the statistical problems discussed in the next subsection.

Predicting Aggregate Returns with Predetermined Observable Variables

The evidence above shows that past returns contain information about expected returns, but they are a noisy signal. A more powerful test uses predetermined explanatory variables that potentially convey more precise information about expected returns. Much recent research documents such predictability using past information. An incomplete list of the variables in these studies includes expected inflation, yield spreads between long-and short-term interest rates and between low-and high-grade bonds, the dividend-to-price ratio, the earnings-to-price ratio, the book-to-price ratio, and the level of consumption relative to income. Importantly, predictability is stronger when the tests use returns measured over longer horizons, with explanatory power rising to levels of 20–40 per cent at two to four year horizons. Unfortunately, the increased explanatory power does not come without econometric problems. First, the number of independent observations decreases with the return horizon. To accommodate, researchers use overlapping observations, but the adjustments for standard errors to account for this perform poorly for the relatively small sample periods used in these tests. Second, most of the variables listed above are highly persistent (in contrast to lagged returns used in autocorrelation tests), and their innovations are correlated with return innovations, resulting in biased test statistics. Despite these shortcomings, the level of statistical significance and the robust nature of the results – across so many different explanatory variables and across so many worldwide equity markets – strongly argue for a predictable component in aggregate returns.

Patterns in Daily Returns Around Weekends

Consider an exchange where trading takes place Monday–Friday. If the process generating stock returns operates continuously, then Monday returns should be three times the returns expected on each of the other days to compensate for a three-day holding period. Call this the calendar-time hypothesis. An alternative is the trading-time hypothesis: returns are generated only during trading periods, and average returns are the same for each of the five trading days in the week. Inconsistent with both hypotheses, stock returns in many countries are negative, on average, on Monday (French 1980). (In Australia, Korea, Japan and Singapore average returns on Tuesday are negative because of time zone differences relative to the US and European markets.)

What causes the weekend effect? That the pattern exists in so many different markets argues persuasively against many institution-specific explanations. Research has shown that the weekend effect cannot be explained by: differences in settlement periods for transactions occurring on different weekdays; measurement error in recorded prices; market maker trading activity; or systematic patterns in investor buying and selling behaviour. That an explanation has been elusive may not be important: in the post-1977 period in the United States and in numerous other markets, the weekend effect has all but disappeared (see Schwert 2003).

Patterns in Returns Around the Turn of the Year

Keim (1983) and others document that 50 per cent of the annual size premium in the United States is

concentrated in the month of January, particularly in the first week of the year. This finding has been reproduced on many equity markets throughout the world. Blume and Stambaugh (1983) subsequently demonstrated that, after an upward bias in average returns for small stocks (related to the magnitude of bid-ask spreads) had been corrected, the size premium is evident only in January.

What explains this phenomenon? Two hypotheses rely on the buying and selling behaviour of market participants to explain the turn-of-the-year size premium. The first hypothesis attributes the effect to year-end tax-related selling by taxable individual investors of stocks that have declined in price (an attribute shared by many small-cap stocks). In such trades the investor realizes a capital loss which can be used to offset realized capital gains, thereby reducing taxable income. There is much evidence that such tax-related trading occurs at the end of the tax year (which in many countries coincides with the end of the calendar year), but a clear link between such trading and stock return behaviour has not been established. A second hypothesis concerns the impact of institutional 'window dressing' at the end of the calendar year – selling off 'loser' stocks that have declined in price (again, typically small-cap stocks) so they don't appear on year-end statements sent to constituent shareholders. Although there is evidence that institutions behave in this fashion, any resulting impact on stock prices is difficult to distinguish from the impact of tax-loss selling. In the end, large bid-ask spreads and high transaction costs for small-cap stocks preclude the profitable exploitation of the short-term return differences between individual small-and large-cap stocks. As a result, the turn-of-the-year size premium continues to be positive in recent years (see Fig. 2a).

Conclusion

Recent research in finance has revealed stock price behaviour that is inconsistent with the predictions of familiar models. The research on time series predictability, as a whole, is convincing evidence that expected returns are not constant through time. There are reasonable business conditions stories that can account for time variation in expected returns. However, some of the temporal patterns in returns – in particular those relating to calendar turning points – are troubling as they defy economic interpretations.

The evidence on cross-sectional anomalies poses a significant challenge to well-established asset pricing paradigms. Yet, despite mounting evidence, there is little consensus on alternative theoretical models. As such, the focus of future research should be on developing such models. Indeed, one of the most significant contributions of this strand of research has been the recognition of potential alternative sources of risk (for example, risk related to financial distress) and of the potential importance of behavioural models. Importantly, researchers must recognize that the existence of this anomalous evidence does not constitute proof that existing paradigms are 'wrong'. There is the issue of data snooping – much of the empirical research on financial market anomalies is predicated on previous research that documented similar findings with the same data. And although many of these effects have persisted for nearly 100 years, this in no way guarantees their persistence in the future. More research is necessary to resolve these issues.

See Also

► Capital Asset Pricing Model

Bibliography

Ball, R. 1978. Anomalies in relationships between securities' yields and yield-surrogates. *Journal of Financial Economics* 6: 103–126.

Banz, R. 1981. The relationship between return and market value of common stock. *Journal of Financial Economics* 9: 3–18.

Basu, S. 1977. Investment performance of common stocks in relation to their price-earnings ratio: A test of the efficient market hypothesis. *Journal of Finance* 32: 663–682.

Blume, M., and R. Stambaugh. 1983. Biases in computed returns: An application to the size effect. *Journal of Financial Economics* 12: 387–404.

Brennan, M. 1970. Taxes, market valuation, and corporate financial policy. *National Tax Journal* 23: 417–427.
Brennan, M., T. Chordia, and A. Subrahmanyam. 1998. Alternative actor specifications security characteristic, and the cross section of stock returns. *Journal of Financial Economics* 49: 345–373.
Carhart, M. 1997. On the persistence in mutual fund performance. *Journal of Finance* 52: 57–82.
Daniel, K., and S. Titman. 1997. Evidence on the characteristics of cross-sectional variation in stock returns. *Journal of Finance* 52: 1–33.
De Bondt, W., and R. Thaler. 1985. Does the stock market overreact? *Journal of Finance* 40: 793–805.
De Bondt, W., and R. Thaler. 1987. Further evidence on investor overreactions and stock market seasonality. *Journal of Finance* 42: 557–581.
Fama, E. 1976. *Foundations of finance*. New York: Basic Books.
Fama, E., and K. French. 1992. The cross section of expected stock returns. *Journal of Finance* 47: 427–466.
Fama, E., and K. French. 1993. Common risk factors in the returns of stocks and bonds. *Journal of Financial Economics* 33: 3–56.
Fisher, L. 1966. Some new stock-market indices. *Journal of Business* 39: 191–225.
French, K. 1980. Stock returns and the weekend effect. *Journal of Financial Economics* 8: 55–69.
Graham, B., and D. Dodd. 1940. *Security analysis: Principles and technique*. New York: McGraw-Hill Book Company, Inc..
Hawawini, G., and D. Keim. 2000. The cross section of common stock returns: A review of the evidence and some new findings. In *Security market imperfections in worldwide equity markets*, ed. D. Keim and W. Ziemba. Cambridge: Cambridge University Press.
Jaffe, J., D. Keim, and R. Westerfield. 1989. Earnings yields, market values and stock returns. *Journal of Finance* 45: 135–148.
Jegadeesh, N., and S. Titman. 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance* 48: 65–92.
Keim, D. 1983. Size-related anomalies and stock return seasonality: Further empirical evidence. *Journal of Financial Economics* 12: 13–32.
Kothari, S., J. Shanken, and R. Sloan. 1995. Another look at the cross-section of expected stock returns. *Journal of Finance* 50: 185–224.
Kuhn, T. 1970. *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Lakonishok, J., A. Schleifer, and R. Vishny. 1994. Contrarian investment, extrapolation and risk. *Journal of Finance* 49: 1541–1578.
Litzenberger, R., and K. Ramaswamy. 1979. The effects of personal taxes and dividends on capital asset prices: Theory and empirical evidence. *Journal of Financial Economics* 7: 163–195.
Lo, A., and C. MacKinlay. 1990. When are contrarian profits due to stock market overreaction? *Review of Financial Studies* 3: 175–205.
Mehra, R., and E. Prescott. 1985. The equity premium: A puzzle. *Journal of Monetary Economics* 15: 145–161.
Miller, M., and M. Scholes. 1982. Dividend and taxes: Some empirical evidence. *Journal of Political Economy* 90: 1118–1141.
Neiderhoffer, V., and M. Osborne. 1966. Market making and reversal on the stock exchange. *Journal of the American Statistical Association* 61: 897–916.
Reinganum, M. 1981. A misspecification of capital asset pricing: Empirical anomalies based on earnings yields and market values. *Journal of Financial Economics* 9: 19–46.
Rosenberg, B., K. Reid, and R. Lanstein. 1985. Persuasive evidence of market inefficiency. *Journal of Portfolio Management* 11: 9–17.
Schwert, G. 2003. Anomalies and market efficiency. In *Handbook of the economics of finance*, ed. G. Constantinides, M. Harris, and R. Stulz, Vol. 1. Amsterdam: North Holland.

Financial Market Contagion

Morgan Kelly

Abstract

The power of the metaphor of contagion – that beliefs, actions, and strategies spread among economic agents like pathogens among biological organisms – causes it to recur in disparate areas of economics. This article focuses on four applications of contagion to economics: social influence or memoryless learning; Bayesian social learning; strategy choice in coordination games; and the spread of crises in international financial markets.

Keywords

Arbitrage; Bayesian learning; Competitive exclusion; Contagion; Coordination games; Ergodic theorem of Markov chains; Financial market contagion; Information cascades; Overconfidence; Popularity weighting; Portfolio allocation; Probability; Social influence

JEL Classifications

G1

Social Influence

The metaphor of contagion is central to the early studies of crowd psychology of Mackay (1841), Tarde (1900) and LeBon (1895); and classical early models of disease diffusion were applied to financial markets by Shiller (1984).

The modern analysis of social influence starts with Allport and Postman (1946–47) who studied the spread of wartime rumour. They identified four circumstances that facilitate the spread of rumour: two are characteristics of the rumour, two of the population. The topic of the rumour should be important to people and the rumour should be hard to verify individually; while individuals should be credulous, and going through a time of unusual stress.

Motivations for neglecting formal Bayesian learning differ between economics and sociology. Sociology emphasizes situations that do not lend themselves to Bayesian updating either through lack of time (is a bank about to fail?), or the nature of the question (what is the one true religion?). Economics, by contrast, emphasizes computational simplicity: rules of thumb make fewer cognitive demands on agents than formal updating algorithms.

Kirman (1993) analyses a simple model of influence that is motivated by the foraging behaviour of ants, but applicable, he argues, to the behaviour of stock market investors. Faced with a choice between two identical piles of food, ants switch periodically from one pile to the other. Kirman supposes that there are N ants and that each switches randomly between piles with probability ε (this prevents the system getting stuck with all at one pile or the other), and imitates a randomly chosen other ant with probability δ.

By the ergodic theorem of Markov chains, there is a unique steady state distribution of ants between piles, and Kirman shows by simulation that the shape of the distribution depends on the relative magnitudes of the imitation parameter δ and the mutation parameter ε. With weak imitation and strong mutation there is a single peak at $\frac{1}{2}$, with equal numbers of ants at each pile. With stronger imitation and weaker mutation, the steady state distribution has two peaks at 0 and N: most ants concentrate on a single pile and switch periodically to the other – the behaviour observed among real ants and possibly stock market participants. In contrast to Bayesian learning models, the absence of martingale convergence allows society continually to flip between beliefs.

The independent work of Weidlich and Haag (1983) in quantitative sociology presents an analogous model in continuous time. Agents switch states with a logistic probability that again depends on the relative social popularity of each choice, but Weidlich and Haag also allow agents to have a personal preference for one of the choices. Again, for sufficiently strong imitative behaviour there is a steady state distribution with two peaks, but now the relative magnitude of the peaks depends on how much agents prefer each choice. Society spends most time at the choice preferred by each agent, but will spend time at the choice that is less popular with everyone, as a consequence of social influence.

Ellison and Fudenberg (1993) look at the role of popularity weighting in choosing between a superior and an inferior technology. They observe that popularity can be a useful summary of the relative past performances of the two technologies – the better technology should be more popular – but that the amount of information conveyed by popularity is diluted the more people rely on it. They therefore look at the likelihood that the better technology is adopted, allowing a fixed fraction of the population to change its choice each period, when the relative weights put on the popularity of the technology versus its performance in the last period are allowed to vary.

Ellison and Fudenberg (1993) show that there is an optimal popularity weighting that causes the system to converge to everyone's using the better technology. If popularity weighting exceeds this optimum, the system converges to a steady state where everyone uses one technology, but which technology depends on the starting number of users of each. With under-weighing of popularity, the inefficient alternative can survive indefinitely.

The competitive exclusion principle, proven in the context of ordinary differential equations by Levin (1970), states that the number of coexisting

species cannot exceed the number of resources they compete for. Here there are two competing species or technologies competing for one resource, being used by people, so if the technological choice problem is recast as one of biological competition we know that only one technology will survive. This is done by Juang (2001), who uses an evolutionary selection argument to show how an Ellison–Fudenberg society can reach the optimum when different groups of agents have sufficiently different popularity weightings. In periods when the inferior technology is excessively popular, agents putting low weight on popularity receive higher payoffs and increase in number, while agents who put high weight on popularity do better in periods when the superior technology is popular.

In the popularity weighting models of Kirman (1993), Weidlich and Haag (1983) and Ellison and Fudenberg (1993), every person is equally influenced by every other member of society. In many situations however, we are influenced more by individuals whom we know and have learned to trust than by strangers. To model the greater social influence of neighbours, the individual is put into some mathematical space, where he or she is more likely to interact with individuals close by than far away. Durlauf (1997) looks at the behaviour of agents in an Ising model (originally developed to model the flipping of magnetic poles of atoms in a crystal) where agents live on a lattice and change between two actions at a rate that depends logistically on the state of their nearest neighbours.

If the influence of neighbours lies above a critical value, the system has two steady state distributions (there are an infinite number of agents so the ergodic theorem of Markov chains does not apply) with all agents either in one state or the other. If agents have a preference for one state over the other (the physical analogue is an external magnetic field) however, the system has only one steady state with all choosing the preferred action.

In Durlauf's model, agents in each state influence each other symmetrically, affecting only their nearest neighbours. Durrett and Levin (1998) analyse a system where agents of different types can affect others over different distances. While biologically motivated – Durrett and Levin (1998) are interested in how slow-growing trees can out-compete rapidly growing grasses – this analysis suggests how propaganda and advertising can be used to cause bad ideas to drive out good ones.

Suppose that type 0 dominates type 1: an agent of type 0 converts a type 1 neighbour at rate 1, whereas a type 1 agent converts a type 0 only at rate $\delta < 1$. If both types have the same radius of influence then, so long as the dominant type 0 avoids getting wiped out by an unlucky run at the start, it will take over. However, Durrett and Levin (1998) show that if the dominant type affects only neighbours in a radius of 1, whereas the dominated type affects neighbours over a large radius R, there is a critical value of the conversion rate $\delta_c < 1$, above which the dominated type 1 takes over.

It is straightforward to demonstrate the existence of social influence empirically when individuals observe the overall popularity in society rather than among neighbours. The influence of best-seller lists on book buying is sufficiently well known for publishers to seek to manipulate them by buying books in stores known to be tracked by the lists, and a variety of examples of imitative behaviour are given by Bikchandani et al. (1992) and Chamley (2004, pp. 59 – 60).

Testing for the influence of neighbours is more difficult because neighbourhood choice is frequently endogenous: one must make sure that the behaviour one is attributing to the influence of neighbours is not due to some individual factor that led the person to choose this neighbourhood over others in the first place.

The classic Ryan and Gross (1943) study, which found that the main factor influencing farmers to adopt hybrid corn was the number of nearby farmers who had adopted it, passes the exogeneity test: it is unlikely that farmers chose farms in order to be near other innovative farmers. Sacerdote (2001) uses the random allocation of roommates to incoming Dartmouth University students to show how roommates influence each others' behaviour, finding that roommates have an effect on individual academic performance, while

dormitory effects influence decisions to join fraternities. Kelly and O Grada (2000) look at the behaviour of Irish immigrants, mostly housemaids and day labourers, in 1850s New York during two bank runs. Since they are immigrants it is possible to identify their social network from their place of origin in their home country: newly arrived immigrants tend to associate with people they knew at home. Kelly and O Grada (2000) found that immigrants from one set of counties in Ireland tended to close their accounts during the panics, while otherwise identical immigrants from other counties stayed put.

Bayesian Learning

Bayesian models of social learning allow individuals to infer the information of other agents from their observed actions in an optimal manner rather than through ad hoc imitation. Bayesian social learning can exhibit pathologies. After the first few agents have chosen, subsequent actions convey little new information and are dominated by idiosyncratic noise. Society converges slowly to the optimal action and, in some circumstances, may become stuck on the suboptimal action. A useful textbook discussion of the literature is given by Chamley (2004).

In Bikchandani et al. (1992) and Banerjee (1992), the world can be in either state σ_0 or σ_1. Each agent receives a signal s_0 or s_1 with symmetric precisions $P(s_0|\ \sigma_0) = P(s_1|\ \sigma_1) = p$ and must choose whether or not to invest. Agents choose in a fixed order and, before receiving his private signal, the agent investing in period t observes the history of past investments and uses this to determine their prior probability π_{1t} that the state is 1.

Bikchandani et al.(1992) start with the case where the cost of investment is $\frac{1}{2}$, the payoff in state 1 is 1, and 0 otherwise. Their expected payoff is $p\pi_{It}/(p\pi_{It} + (1-p)(1-\pi_{It}))$. After a number of moves there will be a sufficient difference between the number who has invested and those who have not for the agent's action to be determined solely by his prior belief, irrespective of his signal. Specifically, if the first agent gets a good signal, the second invests if he gets a good signal, and all subsequent agents will then invest irrespective of their signals. If the second gets a bad signal he is indifferent about investing and is assumed to invest, so the third investor again invests regardless of signal, and so on. Once there are two more investors than non-investors, the excess of positive signals outweighs any negative signal an agent might have. Everyone invests regardless of signal, leading to a cascade.

An unlucky series of wrong signals at the start of the game can lead society to fix on the wrong equilibrium. Bikchandani et al. (1992) observe that this wrong equilibrium is fragile, being based on the observations of a handful of early agents, and vulnerable to being overturned by public information available to all agents.

A frequent criticism of cascade models is their reliance on finite signals: all signals are equal and there is no way for a huge negative signal to counteract a series of positive ones. However, the important lesson of the cascade literature is not that society can get stuck at the wrong equilibrium – which requires signals that are finite – but that Bayesian learning when individual signals are observed imperfectly is very slow to converge to the true equilibrium. Vives (1993) shows how adding noise to a Gaussian model slows down its convergence from rate t to rate $t^{1/3}$: 1,000 noisy observations are equivalent to ten clean ones.

The basic intuition of cascades models that imperfectly observed individual information is poorly incorporated into social beliefs is the basis of several other models. Bulow and Klemperer (1994) model rational frenzies in auctions where participants reveal their valuations by bidding. Bidders with high valuations are willing to pay just under the Walrasian clearing price and, being usually inframarginal, all face similar optimization problems. A bid by one agent therefore sets off a chain of bidding by other agents, leading to a pattern of booms and crashes. Caplin and Leahy (1994) look at investment where individuals have Gaussian signals. If the true state is bad, individuals continue to invest, driven by the dominating effect of past actions. Eventually, however, because signals are Gaussian, a few agents get sufficiently bad signals to induce them to stop

investing, causing priors rapidly to move to a belief that the state is bad, leading to a market crash and 'wisdom after the fact'.

While the essence of the cascade literature is that agents transmit a noisy signal of their information, Avery and Zemsky (1998) observe that this is not the case for markets obeying the efficient markets hypothesis where price reflects all publicly available information. In such markets, assuming risk-neutral agents, the price of an asset worth 1 in the good state and 0 in the bad is the Bayesian prior π_{1t}, causing agents always to trade according to their private signal. They show that cascades can still occur if extra dimensions of uncertainty are added – specifically if there is event uncertainty (agents know that something important has happened by whether it is good or bad), or compositional uncertainty (agents are uncertain how many informed traders are active in the market).

Underlying Bayesian models of cascades is the obvious but strong assumption that people are Bayesians. Probability is difficult for most people, and conditional probability especially so. Even with trivial problems of the form 'a family has two children, one of whom is a daughter: what is the probability that the other child is a son?' most will incorrectly answer $\frac{1}{2}$ rather than $\frac{2}{3}$. Similarly, when asked 'one per cent of the population has a disease. A test detects the disease in 95 per cent of patients when it is present, and generates ten per cent false positives when it is absent. What is the probability that someone who tests positive has the disease?', most will give answers slightly below 95 per cent rather than the correct 1.05 per cent.

In other words, people appear to ignore base rates, assuming that the probability of a state given a signal equals the probability of the signal given the state $P(\sigma_i|\ s_i) = P(s_i|\ \sigma_i)$ even when the probability of the state is considerably lower than the probability of the signal. Agents show overconfidence, focusing excessively on their own signal rather than the history of signals of other agents contained in the prior.

If people neglect priors in this way, cascades cannot occur when private signals are uncorrelated. However, if the signal is common, cascades can still occur. For instance if agents view market price as the signal, a run of rising prices induced by improving fundamentals (such as the good macroeconomic conditions and loose credit that Kindleberger (1978) saw as the preconditions for speculative bubbles) are treated by agents as a positive signal inducing them to buy, driving up price and inducing others to buy, and so on.

Strategies in Coordination Games

Kandori et al.(1993) considered the strategies of players in a coordination game with payoffs

	L	R
L	a, a	b, c
R	c, b	d, d

where $a > c$, $d > b$ and $(a - c) > (d - b)$ so (L, L) and (R, R) are Nash equilibria and $(L,$ is the risk dominant one. With myopic, best-response play, they show that a small probability of mutation suffices for the risk dominant equilibrium to be chosen.

Ellison (1993) observed that this convergence is slow, requiring many simultaneous mutations, and showed instead that if there is local interaction of players along a line, the $\frac{1}{2}$ -dominant strategy (the best response if half your neighbours adopt it) spreads rapidly, but not in two dimensions. Blume (1995) shows that non-trivial mixed long run equilibria exist in two dimensional interaction but not in one, while Morris (2000) examines the characteristics of arbitrary networks that permit the risk-dominant strategy to spread. Lee and Valentinyi (2000) look at a game without mutation but where initial strategy choice is random and show that myopic best response to strategies played by immediate neighbours on the lattice causes large populations to coordinate on the risk dominant equilibrium.

International Market Contagion

Large falls in asset values in one country are sometimes followed rapidly by falls in other

countries. To the extent that these falls are too great to be explained by interdependence in trade or exposure to common macroeconomic factors, the process is called contagion.

Two main sources of contagion have been proposed: financial fragility and common financial linkages; and pathologies in the diffusion of information. The empirical study of Kaminsky, Reinhart and Vegh (2003) argues that three sources of fragility underlie international contagion: rapid inflows of capital; macroeconomic shocks that occur too rapidly for gradual portfolio rebalancing; and a leveraged common creditor. Allen and Gale (2000) show that if banks in different regions have claims on each other, a fall in asset values in one region can bring banks in other regions under pressure and lead to falls in asset values in those regions. In Kyle and Xiong (2001) losses suffered by traders who arbitrage between markets dominated by fundamentalists and markets dominated by noise traders cause traders to reduce their positions in both markets, leading to returns becoming more volatile and more correlated.

Models of contagion as information transmission abstract away from agents who revise excessively optimistic forecasts of returns in all markets after a fall in one market, and concentrate on rational actors instead. Calvo and Mendoza (2000) show that if there are fixed costs to gathering and processing information specific to one country and limits to short selling in each country, the benefits of acquiring information about each country in one's portfolio fall as the portfolio expands. Agents put more weight on the behaviour of other investors, making portfolio allocation more sensitive to realized returns in each market. In Kodres and Pritsker (2002), portfolio rebalancing by informed investors can set off panics among the uninformed who misinterpret it as negative information about the market.

The empirical literature on testing for contagion has focused on increases in the correlation of returns between markets during periods of crisis. Forbes and Rigobon (2002) show the elementary weakness of simple correlation tests: with an unchanged regression coefficient, a rise in the variance of the explanatory variable reduces the coefficient standard error, causing a rise in the correlation of a regression.

The regression underlying contagion tests is of the form

$$y_{it} = \delta_i' \mathbf{z}_t + \alpha_i' \mathbf{x}_{it} + \beta_i I\left(y_j - c_j\right) + \varepsilon_{it}$$

where y_i is asset return in country i, z are common macroeconomic factors, x_i are country specific factors, and I is an indicator of a period of crisis in the originating economy j. As Pesaran and Pick (2007) observe, this is a difficult system to estimate econometrically. To disentangle contagion from interaction effects, county-specific variables have to be used to instrument foreign returns. Choosing the crisis period introduces sample selection bias, and it has to be assumed that crisis periods are sufficiently long to allow correlations to be reliably estimated. In consequence, there appears to be no strong consensus in the empirical literature as to whether contagion occurs between markets, or how strong it is.

See Also

► Information Cascades

Bibliography

Allen, F., and D.M. Gale. 2000. Financial contagion. *Journal of Political Economy* 108: 1–33.

Allport, G.W., and L. Postman. 1946–47. An analysis of rumor. *Public Opinion Quarterly* 10: 501–517.

Avery, C., and P. Zemsky. 1998. Multidimensional uncertainty and herd behavior in financial markets. *American Economic Review* 88: 724–748.

Banerjee, A.V. 1992. A simple model of herd behaviour. *Quarterly Journal of Economics* 107: 797–817.

Bikchandani, S., D. Hirshleifer, and I. Welch. 1992. A theory of fads, custom, and cultural change as informational cascades. *Journal of Political Economy* 100: 992–1026.

Blume, L. 1995. The statisticial mechanics of best-response strategy revision. *Games and Economic Behavior* 11: 111–145.

Bulow, J., and P. Klemperer. 1994. Rational frenzies and crashes. *Journal of Political Economy* 102: 1–23.

Calvo, G.A., and E.G. Mendoza. 2000. Rational contagion and the globalization of security markets. *Journal of International Economics* 51: 79–113.

F

Caplin, A., and J. Leahy. 1994. Business as usual, market crashes, and wisdom after the fact. *American Economic Review* 84: 548–565.
Chamley, C. 2004. *Rational herds: Economic models of social learning*. Cambridge: Cambridge University Press.
Durlauf, S.N. 1997. Statistical mechanics approaches to socioeconomic behavior. In *The economy as an evolving complex system II*, ed. B. Arthur, S. Durlauf, and D. Lane. Reading, MA: Addison-Wesley.
Durrett, R., and S. Levin. 1998. Spatial aspects of interspecific competition. *Theoretical Population Biology* 53: 30–43.
Ellison, G. 1993. Learning, local interaction, and coordination. *Econometrica* 61: 1047–1071.
Ellison, G., and D. Fudenberg. 1993. Rules of thumb for social learning. *Journal of Political Economy* 101: 612–643.
Forbes, K.J., and R. Rigobon. 2002. No contagion, only interdependence: Measuring stock market comovements. *Journal of Finance* 57: 2223–2261.
Juang, W.-T. 2001. Learning from popularity. *Econometrica* 69: 735–747.
Kaminsky, G.L., C.M. Reinhart, and C.A. Vegh. 2003. The unholy trinity of financial contagion. *Journal of Economic Perspectives* 17(4): 51–74.
Kandori, M., G.J. Mailath, and R. Rob. 1993. Learning, mutation, and long run equilibria in games. *Econometrica* 61: 29–56.
Kelly, M., and C. O Grada. 2000. Market contagion: Evidence from the panics of 1854 and 1857. *American Economic Review* 90: 1110–1125.
Kindleberger, C. 1978. *Manias, panics, and crashes*. New York: Basic Books.
Kirman, A. 1993. Ants, rationality, and recruitment. *Quarterly Journal of Economics* 108: 137–156.
Kodres, L., and M. Pritsker. 2002. A rational expectations model of financial contagion. *Journal of Finance* 57: 769–800.
Kyle, A., and W. Xiong. 2001. Contagion as a wealth effect. *Journal of Finance* 56: 1401–1440.
LeBon, G. 1895. *Psychologie des Foules*. Translated as *The Crowd: A Study of the Popular Mind*, New York: Viking, 1960.
Lee, I.H., and A. Valentinyi. 2000. Noisy contagion without mutation. *Review of Economic Studies* 67: 47–56.
Levin, S.A. 1970. Community equilibria and stability, and an extension of the competitive exclusion principle. *American Naturalist* 104: 413–423.
Mackay, C. 1841. *Extraordinary popular delusions and the madness of crowds*. London: Bentley.
Morris, S. 2000. Contagion. *Review of Economic Studies* 67: 57–78.
Pesaran, M.H., and A. Pick. 2007. Econometric issues in the analysis of contagion. *Journal of Economic Dynamics and Control* 31: 1245–1277.
Ryan, B., and N.C. Gross. 1943. The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology* 8: 15–24.
Sacerdote, B. 2001. Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics* 116: 681–704.
Shiller, R.J. 1984. Stock prices and social dynamics. *Brookings Papers on Economic Activity* 1984(2): 457–498.
Tarde, G. 1900. *Les Lois de l'Imitation*. Translated as *The Laws of Imitation*, Gloucester, MA: Peter Smith, 1962.
Vives, X. 1993. How fast do agents learn? *Review of Economic Studies* 60: 329–347.
Weidlich, W., and G. Haag. 1983. *Concepts and models of quantitative sociology*. New York: Springer.

Financial Markets

Nils H. Hakansson

One of the more noteworthy developments in economics over the last twenty years or so is the emergence of equilibrium models of the financial market. Included in this term is the market for financial securities such as stocks, bonds, options and insurance contracts. The chief building block and spur in this evolution has been the economics of uncertainty, which itself is of rather recent origin. The results of this new focus and the activities and synergies it has generated is often broadly referred to as financial economics. It is within this new subfield that various models of the financial market occupy the centre stage.

After a brief summary of models based on analysis in return space in section, "Return Space Analysis", this essay will focus on the two-period, pure-exchange model of the financial market beginning in section, "The Basic Model". Conditions under which full efficiency is attained in incomplete markets will be identified in section, "Full Allocational Efficiency in Incomplete Markets". Finally, section, "Changes in the Financial Market" will trace the welfare and price effects resulting from changes in the financial market.

Return Space Analysis

Much of the earlier work in financial equilibrium focused on pay-off returns rather than total pay-offs

or consumption levels. While return space is both a natural and intuitive object of concern, and in fact continues to draw much attention, it faces certain shortcomings in addressing many questions of interest where prices, endowments and consumption pay-offs play a central role. I shall therefore provide only a brief review of the main results in return space before moving on to the consumption- and wealth-oriented models.

The so-called capital asset pricing model (CAPM) was more or less independently developed by Sharpe (1964)), Lintner (1965), and Mossin (1966). It studies a single-period, frictionless, competitive market of financial securities. Assuming that (a) investors' preferences are a function of only the mean and the variance of the portfolio's anticipated return (with the mean favoured and the variance disfavoured), (b) investors have homogeneous probability assessments of returns and (c) there is a risk-free asset and that unlimited borrowing is available at the lending rate, three principal results are obtained in equilibrium:

1. The expected return on an optimal portfolio is a positive linear function of its standard deviation of return.
2. The expected return on every security (and portfolio) is a positive, linear function of its (return) covariance with the market portfolio of risky assets (the portfolio which includes x per cent of the outstanding shares of all securities in the market).
3. All optimal portfolios are comprised of the market portfolio of risky assets in conjunction with either risk-free borrowing or lending.

Since the CAPM model is consistent with the von Neumann and Morgenstern (1944) theory of rational choice only under quadratic preferences and/or normally distributed returns, it has left many economists uncomfortable. Nevertheless, it has been the basis of a very large number of empirical studies, which on balance show that the CAPM model provides a rather good first approximation of observed return structures in the financial markets of various countries.

A more recent development is the so-called arbitrage pricing theory (APT) developed by Ross (1976). It posits that security returns are generated by a linear K-factor mode (with K small) in which securities' residual risks are sufficiently independent across securities for the law of large numbers to apply. APT can therefore be viewed as an extension of the single-index model introduced by Markowitz (1952) and developed and extended by Sharpe (1963; 1967), which in turn, of course, is closely related to the CAPM. Not surprisingly, the APT appears to offer a somewhat better fit than the CAPM or single-index model.

In studying the economics of financial markets, however, the CAPM and the APT frameworks do not offer fertile ground. In the CAPM framework, for example, the capital structures of firms are a matter of indifference. To study these and other questions, we must therefore turn to more comprehensive formulations.

F

The Basic Model

The earliest models systematically incorporating uncertainty in analysing markets were those of Allais (1953), Arrow (1953), Debreu (1959, Ch. 7) and Borch (1962). They may therefore be viewed as the forerunners of more comprehensive models of the financial market, including the two-period model developed below.

Assumptions

We consider a pure-exchange economy with a single commodity which lasts for two periods under the standard assumptions. That is, at the end of period 1 the economy will be in some state s where $s = 1,\ldots, n$. There are I consumer-investors indexed by i, whose probability beliefs over the states are given by the vectors $\pi_i = (\pi_{i1},\ldots, \pi_{in})$, where, for simplicity, $\pi_{is} > 0$, all i, s.. The preferences of consumer-investor i are represented by the (conditional) functions $U_{is}(c_i, w_{is})$, where c_i is the consumption level in period 1 and w_{is} is the consumption level in period 2 if the economy is in state s at the beginning of that period. These functions are defined for

$$(c_i, w_{is}) \geqq 0 \quad \text{all } i, s \tag{1}$$

and are assumed to be increasing and strictly concave.

At the beginning of period 1 (time 0), consumer-investors allocate their resources among current consumption c_i and a portfolio chosen from a set J of securities indexed by j. Security j pays $a_{js} \geq 0$ per share at the end of period 1 and the total number of outstanding shares is Z_j. Let z_{ij} denote the number of shares of security j purchased by investor i at time 0; his portfolio $Z_i = (z_{i1}, \ldots, z_{ij})$ then yields the pay-off

$$w_{is} = \sum_{j \in J} z_{ij} a_{js},$$

available for consumption in period 2 if state s occurs at the end of period 1. Investor endowments are denoted (C_i, Z_i) and aggregate wealth or consumption in state s is given by

$$W_s = \sum_{j \in J} Z_j a_{js}, \quad \text{all } s.$$

The financial markets, as is usual, are assumed to be competitive and perfect; that is, consumer-investors perceive prices as beyond their influence, there are no transaction costs or taxes, securities and commodities are perfectly divisible, and the full proceeds from short sales (negative holdings) can be invested. The number of securities, however, need not be large (although this is not ruled out). Since our focus is on the structure of the financial market, and changes therein, production decisions (and hence the vector of aggregate consumption (C,W)) are viewed as fixed.

If the rank of matrix $A = [a_{js}]$ is full (equals n), the financial market will be called *complete;* if not, it will be called *incomplete*. The significance of a complete market is that *any* pay-off pattern $w \geqq 0$ can be obtained via some portfolio z since the system $zA = w$ will always have a solution. (In incomplete markets, in contrast, some pay-offs patterns $w \geqq 0$ are infeasible). The simplest form of a complete market is that in which $A = I$ (the identity matrix); the financial market is now said to be composed of *Arrow-Debreu or primitive* securities (as opposed to *complex* securities.) The main 'advantage' of an Arrow–Debreu market is that it never requires the consumer-investor to take short positions, which is generally necessary in a complete market composed of complex securities. Finally, a financial market which contains a risk-free asset, or makes it possible to construct a risk-free portfolio, is called *zero-risk compatible*.

Under our assumptions, each consumer-investor i maximizes

$$u_i \equiv \sum_s \pi_{is} U_{is}\left(c_i, \sum_{j \in J} z_{ij} a_{js}\right) \tag{2}$$

with respect to the decision vector (c_i, z_i), subject to (1) and to the budget constraint

$$c_i p_0 \sum_{j \in J} z_{ij} p_j = \bar{c}_i p_0 \sum_{j \in J} \bar{z}_{ij} p_j$$

as a price-taker, where P_0 is the price of a unit of period 1 consumption and P_j is the price of security j.

Equilibria and Their Properties In view of our assumptions, an equilibrium will exist but need not be unique (see e.g. Hart 1974; note also that uniqueness is with reference to the consumption allocation (c, w), not allocation (c, z)). The equilibrium conditions for any market structure A, assuming for simplicity that the non-negativity constrains on consumption are not binding may be written

$$\sum_s \pi_{is} \frac{\partial U_{is}\left(c_i \sum_{j \in J} z_{ij} a_{js}\right)}{\partial c_i} \lambda_i \quad \text{all } i \tag{3}$$

$$\sum_s \pi_{is} \frac{\partial U_{is}\left(c_i \sum_{j \in J} z_{ij} a_{js}\right) a_{js}}{\partial w_{is}} = \lambda_i p_j \quad \text{all } i, j \tag{4}$$

$$(c_i, z_i A) \geq 0 \quad \text{all } i \tag{5}$$

$$c_i + z_i p = \bar{c}_i + \bar{z}_i p \quad \text{all } i \tag{6}$$

$$\sum_i (c_i, z_i) = (C, Z) \tag{7}$$

where the λ_i are Lagrange multipliers, (7) represents the market clearing equations, and P_0 has been chosen as numeraire, i.e. $P_0 \equiv 1$.

Any allocation (c, z) which constitutes a solution to system (3), (4), (5), (6), and (7) (along with a price vector P and a vector λ) is *allocationally efficient with respect to the market structure A* – since the marginal rates of substitution for any two *securities* are the same across individuals. When *(c,z)* is allocationally efficient with respect to *all* conceivable allocations, whether achieved outside the existing market or not, (c, z) will be said to be *fully allocationally efficient* (FAE).

To be more precise, define the *shadow prices* R'_{is} is by

$$R'_{is} \equiv \frac{1}{\lambda}\left(\pi_{is}\frac{\partial U_{is}\left(c_i, \sum_{j\in J} z_{ij}a_{js}\right)}{\partial w_{is}}\right).$$

It is well known that (3), (4), (5), (6), and (7) plus

$$R'_{is} = R'_{1s} \quad \text{all } i \geq 2, \quad \text{all } s \tag{8}$$

is a necessary and sufficient condition for the market allocation (c, z) to be FAE because (8) insures that the marginal rates of substitution of wealth between any two *states* are the same for all investors i. (4) may now be written

$$AR'_{is} = p, \quad \text{all } i. \tag{4'}$$

Implicit Prices

The equilibrium value of a feasible second-period pay-off vector w will be denoted $V(w)$; thus if w is obtainable via portfolio z, we get $w = zA$ and hence

$$V(w) = V(zA) = zp = wR = zAR.$$

In the above expression, $R = (R_1, \ldots, R_n)$ represents the not necessarily unique set of *implicit prices* of (second-period) consumption in the various states implied by P since

$$AR = p. \tag{9}$$

By Farkas' Lemma, a positive implicit price vector is always present in the absence of arbitrage and hence in equilibrium. (Arbitrage is the opportunity to obtain either a pay-off $w \geq 0, w \neq 0$, at a cost $zP \leq 0$, or a pay-off $w = 0$ at a cost $zP < 0$). In view of (4′) and (9), shadow prices are always implicit prices, but a set of implicit prices need not be anyone's shadow prices.

Full Allocational Efficiency in Incomplete Markets

When the financial market A is complete, systems (4′) and (9) have only one solution, which insures that

$$R'_t = R, \quad \text{all } i.$$

This condition, as noted, is necessary and sufficient to attain FAE. Complete financial markets, while a useful abstraction, are not an everyday occurrence, however. Securities number at most a few thousand, while the relevant set of states is no doubt much larger. This leads us to the question: under what circumstances is FAE attained in incomplete markets? One such case is trivial and will be dismissed quickly: the case when individuals are identical in their preferences, beliefs and (the value of their) endowments. We now turn to three other sets of conditions when this occurs.

Diverse Endowments

Are there any conditions under which individuals with *diverse* endowments are as well served by a single security in the market as by many? The answer is yes; beliefs must be homogeneous and preferences e.g. of the form

$$U_{is}(c_i, w_{is}) = \begin{cases} U_i^1(c_i) + \rho_s U_i^2(w_{is}) \\ \text{or} \\ U_i^1(c_i)\rho_s U_i^2(w_{is}) \end{cases} \quad \text{all } i, s \tag{10}$$

(with $\rho_s > 0$), where

$$U_i^2(w_{is}) = (1/\gamma)w_{is}^{\gamma}, \quad \gamma < 1, \quad \text{all } i$$

That is, preferences for second-period consumption must be separable, isoelastic and homogeneous. Everyone's optimal portfolio is now of the form

$$z_i = k_i Z, \quad \text{all } i$$

where the k_i are fractions. In addition, the equilibrium implicit prices R are now unique and completely independent of the market structure A.

Linear Risk Tolerance

To attain FAE with heterogeneous second-period preferences, we need at least two securities in the market. Two-fund separation occurs in every zero-risk compatible market A under homogeneous beliefs (but arbitrary return structures) when preferences are of the form (10) if and only if

$$U_i^2(w_{is}) = \begin{cases} (1/\gamma)(\phi_i + w_{is})^\gamma & \text{all } i \\ \text{or} & \\ -(\phi_i - w_{is})^\gamma \ \gamma > 1, \ \phi_i \text{ large}, & \text{all } i \\ \text{or} & \\ -\exp\{\phi_i w_{is}\} \quad \phi_i < 0, & \text{all } i \end{cases}$$

provided none of the non-negativity constraints on consumption is binding. The optimal policies are now of the form

$$z_i = k_{i1} z' + k_{i2} z'', \qquad \text{all } i,$$

where the portfolio (fund) z' is risk-free and portfolio z'' is risky (see e.g. Rubinstein 1974). It is evident that with diverse endowments, preferences must belong to a very narrow family, even when beliefs are homogeneous, in order for FAE to be attained.

Supershares

Two states s and s' such that $W_s = W_{s'}$, i.e., with equal aggregate pay-offs, are said to belong to the same superstate t (Hakansson 1977). If the financial market is complete with respect to the superstate partition T, FAE is attained for arbitrary endowments if and only if

$$\pi_{is}/\pi_{it} = \pi_{1s}/\pi_{1t}, \quad \text{all } s \in t, \quad \text{all } i \text{ and } t \quad (11)$$

and

$$U_{is} = U_{is'}, \quad \text{all } s \text{ and } s' \in t, \quad \text{all } i \text{ and } t \quad (12)$$

Note that (10) and (11) require only conditionally homogeneous beliefs and that preferences are insensitive to states *within* a superstate -beliefs and preferences with respect to superstates are unrestricted.

To complete the market with respect to superstates, three simple alternatives are available (Hakansson 1978). The first is a full set of 'supershares', each share paying \$1 if and only if a given superstate occurs (superstates are readily denominated in either nominal or real terms). The second and third alternatives are a full set of (European) call options or a full set of (European) put options on the market portfolio αZ or αW, where $0 < \alpha \leq 1$.

It may be noted that a market in puts and calls on a crude approximation to the United States market portfolio, namely the Standard & Poors 100 Index, was opened in 1983. These options are now the most actively traded of all option instruments.

Changes in the Financial Market

Changes in the set of securities available in the financial market are everyday occurrences. Early studies on this subject include those of Borch (1968, Ch. 8), Ross (1976) and Litzenberger and Sosin (1977). To trace fully the effects of such changes involves comparing equilibria, which is a matter of some complexity. However, using the two-period framework of this assay, it is possible to reach some general conclusions on how changes in the market structure from A' to A'', say, affect welfare, prices and other dimensions of interest in a pure exchange setting.

The Feasible Allocations

One of the critical determinants, not surprisingly, is the change in feasible allocations. Recall that a market structure A is any 'full' set of instruments; that is, any set of instruments capable of

$$F(A) \equiv \left\{ w \,\middle|\, w_i \geqq 0, w_i = z_i A, \sum_t z_{ij} = Z_j, \quad \text{all } j \right\}.$$

allocating, in some fashion, aggregate wealth $W = (W_1, \ldots, W_n)$. The set of feasible second-period consumption allocations $w = (w_1, \ldots, w_1)$ obtainable via market structure A will be denoted $F(A)$, i.e.

In comparing two market structures A' and A'' with respect to feasible allocations, there are (since holding the market portfolio αZ is always feasible) three possibilities; either

$$F(A') = F(A'') \qquad \text{(Type I)}$$

or

$$F(A') \subset F(A'') \,(\text{or the converse}) \qquad \text{(Type II)}$$

or

$$\{F(A') \cap F(A'')\} \subset F(A')$$
$$\{F(A') \cap F(A'')\} \subset F(A''). \qquad \text{(Type III)}$$

These three types of changes will be referred to as feasibility preserving, feasibility expanding (or reducing) and feasibility altering.

A sure way to obtain a feasibility expanding change is to make a finer and finer breakdown of *existing* instruments into an ever larger set of linearly independent (or unique) securities.

Endowment Effects

Since changes in the financial market structure are generally implemented by firms or exchanges and take place when the market is closed, such changes frequently alter investors' endowments. An example would be a merger, which results in the substitution of new securities for old. It is useful to distinguish between three cases:

1. *Strong Endowment Neutrality This occurs if the endowed consumption patterns in the two markets are unaltered, i.e. if*

$$\left(\overline{c}_i', \overline{w}_i'\right) = \left(\overline{c}_i'', \overline{w}_i''\right), \quad \text{all } i.$$

2. *Weak Endowment Neutrality* This occurs if the values of the endowments (provided there is a common implicit equilibrium price structure R) are identical in the two markets, i.e. if

$$\begin{aligned} \overline{c}_i' + \overline{z}_i' p' &= \overline{c}_i' + \overline{w}_i' R + \overline{c}_i'' + \overline{w}_i'' R \\ &= \overline{c}_i'' + \overline{z}_i'' p'', \quad \text{all } i \end{aligned}$$

where $R > 0$ satisfies $A'R = P'$ and $A''R = P''$.

3. *Non-Neutral Endowment Changes* While the first two cases are rather rare, strong endowment neutrality typically accompanies non-synergistic *(pro rata)* corporate spin-offs when applicable bonds remain risk-free, as well as the opening of option markets, for example.

The Welfare Dimension

As noted, in comparing different market structures, the comparison which is ultimately relevant is that which compares allocations actually attained; that is, equilibriumn allocations. Using (2), we denote investor i's equilibrium expected utility in market all i structure A'' by u_i'' and his equilbrium expected utility in market structure A' by u_i'. A comparison of any given equilibrium in market A'' with some equilibrium in some other market A' must then yield one of four cases:

$$u_i'' \geq u_i', \text{ all } i, \; u_i'' > u_i', \text{ some } i \,(\text{Pareto dominance}) \qquad \text{(i)}$$

or

$$u_i'' = u_i', \text{ all } i \qquad (\text{Pareto dominance}) \qquad \text{(ii)}$$

or

$$u_i'' > u_i', \text{ some } i, \; u_i'' > u_i', \text{ some } i \,(\text{Pareto redistribution}) \qquad \text{(iii)}$$

or

$$u_i'' \leq u_i', \text{ all } i, \; u_i'' > u_i', \text{ some } i \,(\text{Pareto inferiority}) \qquad \text{(iv)}$$

The task at hand, then, is to identify the conditions under which each of these cases, as well as

F

combinations of these cases, will occur. All comparisons are contemporaneous in the sense that they compare welfare under market structure A'' to what it would be if Á′ were in use instead.

Principal Results

The principal results (Hakansson 1982) may be summarized as follows:

1. Feasibility preserving market structure changes yield either Pareto equivalence or redistributions. To preclude Pareto redistributions we must either have efficient endowments in the first market and strong endowment neutrality, or weak endowment neutrality coupled with unique equilibria. Pareto equivalence is always accompanied by value conservation.
2. Feasibility expanding market structure changes imply either Pareto dominance, Pareto equivalence or Pareto redistributions. To preclude redistributions we must have efficient endowments in the first market and strong endowment neutrality, or weak endowment neutrality coupled with unique equilibria. Value conservation is highly unlikely.
3. Feasibility altering changes in the market structure have unpredictable value and welfare effects.
4. Value and welfare effects are relatively independent.

As noted by Hart (1975), the introduction of multiple commodities or more than two periods is a non-trivial step which may bring about additional complications, such as Pareto-dominated equilibria when feasibility is expanded.

Within the limits of the single-good, two-period model under pure exchange, certain tentative general conclusions concerning common market structure changes can be stated. Even under mild heterogeneity of preferences and/or beliefs, 100 per cent non-synergistic mergers tend to be welfare reducing while (non-synergistic) spin-offs and the opening of option markets tend to be beneficial. The use of risky bonds and preferred stock tends to be virtuous as well, at least apart from bankruptcy costs. Finally, value conservation is a much rarer phenomenon than suggested by Modigliani and Miller (1958) and Nielsen (1978) among others.

See Also

- ▶ Capital, Credit and Money Markets
- ▶ Capital Asset Pricing Model
- ▶ Finance

Bibliography

Allais, M. 1953. L'extension des théories de l'équilibre économique général et du rendement social au cas du risque. *Econometrica* 21: 269–290.

Arrow, K. 1953. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies* 31: 91–96, 1964.

Borch, K. 1962. Equilibrium in a reinsurance market. *Econometrica* 30: 424–444.

Borch, K. 1968. *The economics of uncertainty*. Princeton: Princeton University Press.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Hakansson, N. 1977. The superfund: Efficient paths toward efficient capital markets in large and small countries. In *Financial decision making under uncertainty*, ed. H. Levy and M. Sarnat. New York: Academic Press.

Hakansson, N. 1978. Welfare aspects of options and supershares. *Journal of Finance* 33(3): 759–776.

Hakansson, N. 1982. Changes in the financial market: Welfare and price effects and the basic theorems of value conservation. *Journal of Finance* 37(4): 977–1004.

Hart, O. 1974. On the existence of equilibrium in a securities model. *Journal of Economic Theory* 9(3): 293–311.

Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11(3): 418–443.

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.

Litzenberger, R., and H. Sosin. 1977. The theory or recapitalizations and the evidence of dual purpose funds. *Journal of Finance* 32(5): 1433–1455.

Markowitz, H. 1952. Portfolio selection. *Journal of Finance* 7: 77–91.

Modigliani, F., and M. Miller. 1958. The cost of capital, corporation finance, and the theory of investment. *American Economic Review* 48: 261–297.

Mossin, J. 1966. Equilibrium in a capital asset market. *Econometrica* 34(4): 768–783.

Nielsen, N. 1978. On the financing and investment decisions of the firm. *Journal of Banking and Finance* 2(1): 79–101.

Ross, S. 1976a. Options and efficiency. *Quarterly Journal of Economics* 90(1): 75–89.
Ross, S. 1976b. The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13(3): 341–360.
Rubinstein, M. 1974. An aggregation theorem for securities markets. *Journal of Financial Economics* 1(3): 225–244.
Sharpe, W. 1963. A simplified model for portfolio analysis. *Management Science* 9: 277–293.
Sharpe, W. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.
Sharpe, W. 1967. Linear programming algorithm for mutual fund portfolio selection. *Management Science*, Series A 13: 499–510.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.

Financial Structure and Economic Development

John Toye

Abstract

Most early development economists neglected the financial aspects of development, often restricting them to domestic taxation, the self-finance of enterprises and the negotiation of foreign credits. In the 1970s, a few economists proposed that private financial intermediation, operating with market-set interest rates, improved incentives to save and the availability of credit, and allocated savings more efficiently between borrowers. Against this, new institutional economists have argued that financial intermediation involves considerable risks since banks find it difficult to acquire skills in risk assessment. The relationship between increases in real income and the size and complexity of the financial superstructure remains loose.

Keywords

Adverse selection; Asymmetric information; Capital controls; Credit rationing; Development banks; Financial intermediaries; Financial intermediation; Financial interrelations ratio; Financial liberalization; Financial repression; Financial structure and economic development; Goldsmith, R.; Interest-rate liberalization; Kalecki, M.; Microcredit; Moneylenders in developing countries; Moral hazard; New institutional economists; Prudential regulation; Reserve ratio requirement; Risk assessment

JEL Classifications

O16

The question of how financial structure relates to economic development departs from a distinction between an economy's infrastructure of real wealth – its physical assets produced by human labour and natural resources – and a set of financial claims that exists side by side with it and is somehow connected with it. This set of claims consists of short-term and long-term loans and credits and equity securities. A second distinction is between two types of issuers and holders of these financial instruments: non-financial institutions, such as governments, business enterprises and households, whose assets are mainly – but not exclusively – held in physical form, and financial institutions, whose assets and liabilities are mainly financial instruments. This second distinction divides the original question into two parts: the link between the real infrastructure and the volume of financial instruments in the economy, and the link between the real infrastructure and the volume of funds held by the financial institutions.

The US economist Raymond W. Goldsmith (1904–88) provided much of the statistical framework and empirical basis for the examination of these questions. In a lifetime of painstaking scholarly labour, he collected data that allowed comparison across countries of key ratios of real and financial assets, and also of how these ratios changed within individual countries over time. His research suggested one – now widely accepted – statistical generalization, that of the *rising financial interrelations ratio.* As expressed by Gurley and Shaw (1967, p. 257), 'during economic development . . . countries usually experience more rapid growth in financial assets than in

national wealth and national products'. The increase, however, does not continue without limit. This process of financial deepening has been experienced by many of the now developed capitalist countries. However, it tends to be most evident in the early and middle phases of their economic development, after which it levels off. The exceptions to this are higher ratios in periods of repressed inflations during and just after major wars. In the United States, little financial deepening has been noticeable since 1950.

It is also clear that this ratio can be influenced by strategic choices in the quest for development. Countries that adopt state-led development strategies, such as the USSR and its eastern European satellites, exhibited smaller ratios than those of countries that relied on private sector growth to drive their economic development.

As one would expect, developing countries have much lower financial interrelations ratios (between 0.6 and 1.0) than do Europe and North America (between 1.0 and 1.5). This is a reflection of the lower degree of monetization of their economies and the relative lack of separation of the functions of saving and investment. The composition of the value of total financial instruments shows a smaller share of financial institutions in the developing countries, for the same reason (Goldsmith 1969, pp. 44–7).

Compositional Changes in Financial Instruments

The process of development is associated with compositional change, as well as growth, in the stock of financial instruments. The start of financial development is the diffusion of fiduciary money into the economy through the banking system. This is followed by the growth of banking deposits, and then the share of banking deposits declines as new types of financial institution proliferate – building societies, mortgage companies, insurance companies and pension funds – providing financial services that are tailored to special needs.

The main thrust of early development economics neglected these financial aspects of development. Until the 1980s, the main focus of analysis was on the real economy, particularly the accumulation of real physical capital, the acquisition of new human skills and the expansion of international trade. When the problem of financing 'real' development was discussed, it was often restricted to the problem of domestic taxation, the self-finance of enterprises and the negotiation of foreign credits. Michal Kalecki, who greatly influenced the early development literature, explicitly argued that the volume of investment is not subject to financial limits. In the Kaleckian view, financial institutions appear, if at all, as pre-existing constraints on production that have to be removed – for example, rural moneylenders (Kalecki 1972, pp. 145–61; FitzGerald 1990, p. 184) – or as publicly established agencies for channelling resources to sectors of the economy that were desired to expand (see Eshag 1983, pp. 186–8 on development banks). Thus financial development was long a secondary consideration, and it was viewed from the perspective of a government deciding which monetary institutions to create and which to destroy.

The McKinnon–Shaw View of Financial Intermediation

Independently of this dominant post-Keynesian approach, Ronald McKinnon, John Gurley and Ed Shaw in the 1970s elaborated a much more positive view of the role of the growth of private financial intermediaries in development. Using the context of an agricultural sector exhibiting strong technological dualism and lumpy investment, they argued that private financial intermediation, operating with market- set interest rates, improved incentives to save and the availability of credit. It did so by spreading risks and transforming the maturity structure of debt in ways more attractive to both savers and borrowers. Their claim was that financial intermediation would provide the benefits of additional savings and the more efficient allocation of those increased savings between borrowers. On this account, the growth of financial intermediation promotes both capital accumulation and the

diffusion of technical progress by spreading risks more widely and in conformity with people's willingness to bear them (Gurley and Shaw 1960, 1967; McKinnon 1973; Shaw 1973).

This positive view of private financial intermediaries has been criticized by post-Keynesians on both theoretical grounds – an incomplete accounting of all the incentive effects of market-determined interest rates – and empirical grounds – the absence of the predicted incentive effects in the savings, investment and interest rate data. However, the most compelling theoretical critique came from the new institutional economists. Rejecting the assumption of perfect information, they showed how various information asymmetries between the knowledge possessed by the private bankers and the knowledge possessed by their clients (savers and investors) generated a radically altered assessment of the potential benefits and dangers of private financial intermediation.

An important conclusion was that, as interest rates rose, the banks' lending portfolios became riskier. This resulted both from adverse selection, as the marginal borrowers are more liable to default, and from moral hazard, as the marginal borrowers are more likely to invest in high-risk projects. Private banks thus have an incentive to continue to lend at less than the market-clearing rate of interest, and to borrow from depositors at an even lower rate, and then to ration credit (Stiglitz and Weiss 1981). The benefits to be expected from private financial intermediation under asymmetric information assumptions are smaller than those derived from McKinnon–Shaw perfect information reasoning.

Evidence of Financial Repression

In the 1980s it became increasingly clear that many existing financial institutions in developing countries were dysfunctional. Moreover, in many cases the cause of the dysfunction was diagnosed as inappropriate government regulation. The analysis of 'financial repression' in Shaw (1973) was often borne out in reality. Interest rates were administered and maximum rates held very low, while reserve ratio requirements were set very high to force banks to buy and hold government debt issued at below-market rates of interest. Banks were treated as a source of government finance, rather than providers of financial services to the private sector. Meanwhile capital controls were in place to stop the flight of private capital seeking better returns abroad.

The consequences of these widespread interventions included banks' inability to offer attractive rates to depositors; an artificially low level of deposits; a shortage of credit; the rationing of available credit; political pressures directing the allocation of credit; low repayment rates; the accumulation of bad debts; and ultimately the effective insolvency of the banks. The flourishing business of rural moneylenders and 'kerb' markets, despite the charging extortionate rates of interest, was simultaneously observed, with puzzlement, complaint or cynicism, as the observer preferred.

The Move to Financial Liberalization

The policy response was that international organizations and national aid donors pressed for the removal of these policy-induced distortions, and the adoption of reforms aimed at financial liberalization. Interest-rate liberalization was adopted as one of the components of structural adjustment programmes in the 1980s and 1990s. Unfortunately, liberalization was not enough by itself to end the effects of financial repression. While deposits did climb as a share of GNP, there was little expansion of credit to the private sector, as many state banks remained in existence, and their habits of directing credit died hard (World Bank 2005, pp. 207–39). Worse still, financial liberalization led to increasingly frequent financial crises. They were the result of increased competition between private banks, increased opportunities of foreign borrowing for all banks and the serious inadequacy of prudential regulation of the banking system.

As the new institutional economists had pointed out, financial intermediation involves considerable risks, and banks find it difficult to acquire skills in risk assessment, especially when

that skill has not been previously salient. Hence, the possibility of miscalculation of risk is ever present. In addition, bank regulation and supervision is itself a risky business, for the by now familiar reason of asymmetric information, and can provoke banking problems as well as prevent them – and can also be vulnerable to corrupt pressures to look the other way. All of this suggests that the building of functional financial sectors is likely to remain a work in progress for the foreseeable future.

Economic and Financial Development: A Loose Reciprocal Relationship

Few would be inclined to deny that there is a rough parallel between economic and financial development, if periods of several decades are the time period under consideration. As real income and wealth increase, so do the size and complexity of the financial superstructure. Yet this is a loose relationship. The *financial intermediation ratio,* the share of financial institutions' assets in the value of all financial assets, is even more loosely tied to the stock of real wealth. Rather, 'it is to a large extent the result of institutional arrangements and savers' preferences' (Goldsmith 1983, pp. 54).

It is difficult, therefore, to interpret the causal significance of these highly aggregative ratios. It is hard to argue that a given volume or composition of financial assets is a sufficient condition for the development of real sectors of the economy – or even a necessary condition, given that rapid growth has sometimes taken place during periods of deliberate financial repression. We are not, however, obliged to retreat to a view of finance as purely passive, accommodating growth that is driven by other means. Financial innovation has at times sparked off virtuous circles of growth in particular sectors and regions. The microcredit movement in Bangladesh (and elsewhere) in response to extortionate rural moneylending is one recent example where a new financial technology, carefully managed, has been the spur to the growth of the incomes and welfare of poor borrowers. However, if building a functioning formal sector of financial intermediaries is arduous and costly, the evolution of financial structure and real economic development may well be mutually determined, with causation flowing in both directions (Greenwood and Jovanovic 1990).

See Also

- ▶ Credit Rationing
- ▶ Financial Liberalization
- ▶ Goldsmith, Raymond William (1904–1988)
- ▶ Microcredit

Bibliography

Eshag, E. 1983. *Fiscal and monetary problems in developing countries*. Cambridge: Cambridge University Press.

FitzGerald, E. 1990. Kalecki on financing development: An approach to the macroeconomics of the semi–-industrialised economy. *Cambridge Journal of Economics* 14: 183–203.

Goldsmith, R. 1969. *Financial structure and development*. New Haven: Yale University Press.

Goldsmith, R. 1983. *The financial development of India, Japan and the United States: A trilateral institutional, statistical and analytic comparison*. New Haven: Yale University Press.

Greenwood, J., and B. Jovanovic. 1990. Financial development, growth and the distribution of income. *Journal of Political Economy* 98: 1076–107.

Gurley, J., and E. Shaw. 1960. *Money in a theory of finance*. Washington, DC: Brookings Institution.

Gurley, J., and E. Shaw. 1967. Financial structure and economic development. *Economic Development and Cultural Change* 15: 257–68.

Kalecki, M. 1972. *Selected essays on the economic growth of the socialist and the mixed economy.* Cambridge: Cambridge University Press.

Levine, R. 1997. Financial development and economic growth: Views and agenda. *Journal of Economic Literature* 35: 688–726.

Levine, R. 2003. More on finance and growth: More finance, more growth? *Federal Reserve Bank of St Louis Review,* July/August: 31–46.

McKinnon, R. 1973. *Money and capital in economic development*. Washington, DC: Brookings Institution.

Shaw, E. 1973. *Financial deepening in economic development*. New York: Oxford University Press.

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

World Bank. 2005. *Economic growth in the 1990s: Learning from a decade of reform*. Washington, DC: World Bank.

Fine Tuning

Francis M. Bator

'Fine Tuning' was Walter Heller's phrase for fiscal and monetary actions by government aimed at countering deviations in aggregate demand – forecast or actual – from some *target* path of output and associated inflation. The idea marked an important change in doctrine. The goal was not merely to smooth out fluctuations, but to track and output-employment/inflation path chosen from the set of attainable paths according to the preferences of the policymaker (see, however, the entry on "► Functional Finance").

Hyperbole aside, advocates of 'tuning' believe that (1) the economy does not adequately tune itself; and (2) we know enough about its dynamic structure – the lags and multipliers – to achieve better results than a policy unresponsive to unwanted movements in aggregate demand, e.g., a regime of fixed money growth and a 'passive' fiscal policy. (To clinch the case, one has to suppose that politicians will not mess things up – that they will not produce worse results than would a policy of 'non-tuning'.)

Both technical premises have drawn sharp attack.

If the Economy Is 'Classical'

New Classical Macroeconomics (NCM) – much in favour during the past fifteen years among young macro theorists – teaches that, if only the macroeconomic managers would stop meddling, the economy would perform about the way the stochastic version of the perfectly competitive, instantly convergent NCM model predicts it will perform: prices and wage rates would keep all markets more or less continuously cleared, and allocation would remain in the neighbourhood of its quasi-efficient Walrasian (moving) equilibrium. If that is so – an empirical question, and not a matter of methodological aesthetics or political preference – attempts by government to manage aggregate demand are at best an irrelevance, or more likely, the principal cause of macroeconomic inefficiency. Business cycles, insofar as they do not reflect feasibly efficient adjustment to changes in endowments, technology and tastes, are caused by capricious fiscal and monetary policies. Private agents make socially erroneous decisions because they are unable to decipher the behaviour of the government.

The money managers in such an NCM economy, at least in the canonical monetarist version of the story, cannot affect *real* economic magnitudes except by acting capriciously. They control the price level and only that, and should concentrate on making it behave. The fiscal managers, in turn, should stick to the neoclassical business of making the budget conform to the preferences of the electorate with respect to income redistribution and the division of output between private use and public services, present and future. As long as the government and the central bank both behave predictably, aggregate demand, total output, and employment will take care of themselves. (The meaning of efficiency in a macro context is problematic. I use the phrase quasi-efficient to allow for some microeconomic distortions, and for the virtual nonexistence of state-contingent futures markets. Quasiefficiency is, of course, relative to given information sets.)

If the Economy Is Keynesian

Suppose, however, that prices and nominal wage rates (or their rates of change) react to excess supply and demand only sluggishly. Real disturbances give rise to cumulative, self-multiplying quantity responses that are both inefficient and slow to dissipate. Even an anticipated nominal event, for example an increase in money supply brought about by a costless airdrop of currency, causes *real* effects. Then, *in principle*, a disturbance-responsive policy could improve matters.

Not so in practice, opponents say. The coefficients (indeed, the equations) of Keynesian models are too unreliable, and the lags are too variable and too long. As a result, an activist policy – even if free of political constraint – is more likely to do harm

than good. As evidence, they cite the poor performance of the US economy during the late 1960s and 1970s. (On one extreme, NCM view, Keynesian models are no good at all. What appears to be quantitative 'structure' in such models is a mirage; it reflects not durable, exploitable regularities but behaviour that is specific to private agents' expectations of government policy. Any anticipated change in policy will cause rational agents to alter their behaviour; the coefficients will shift the way the Phillips wage-inflation/unemployment relationship shifted in response to the government's attempt during 1962–8 to exploit it. On still another view, Keynesian econometric methodology is inefficient in identifying the economy's true structure. Autoregressive methods that infer structural relations among the variables entirely from the evolving pattern of leads and lags, and make no use of prior theory, are, it is alleged, more likely to reveal robust regularities.)

Pro-activists are quick to acknowledge that Keynesian econometric regularities are approximate and impermanent, and that large shifts in policy regimes may cause them to change. But they read the evidence to say that such 'structural' change is apt to be episodic or gradual or both – that the coefficients are durable enough to be *cautiously* usable. They favour large policy actions only when the gap between aggregate demand and its target is already large, or when the odds are good that it is about to become large. Against small gaps or small disturbances, they would take only small actions or none. Even then, they say, mistakes will occur. But they emphasize how singular the structure of the economy would have to be, and how special the pattern of disturbances, to justify reliance on a 'passive' policy (e.g. trying to keep the various measures of money supply growing at constant rates, and the fiscal instruments fixed in their neoclassically warranted baseline settings).

The 1965–81 US Evidence

Opponents of an activist policy make much of the American experience between 1965 and 1981. But the lesson to be learned from that experience depends critically on whether the US economy is classical or Keynesian. If in fact the economy is Keynesian, then the 1965–81 history provides little or no support for the opponents' case.

In the United States, the acceleration of inflation during 1965–8 was caused not by an overresponsive policy, but by exactly the opposite – the government's failure to heed Keynesian pleas that it counter the excessive thrust of aggregate demand by increasing taxes and making money tight. Plausibly, also, it was that failure, and the resulting rise in the pace of inflation experienced by employers and employees, that caused the Phillips unemployment/wage-inflation regularity of 1946–65 to come unstuck (thus validating the Phelps/Friedman accelerationist prediction, though not necessarily its narrowly expectations-based rationale). That the excess demand of 1965–8 was caused by a large increase in government spending, and not by an unforeseen shift in private spending propensities, made the error of non-tuning the more egregious.

To blame activist policy for the spurts of rapid inflation during the 1970s, or for the simultaneous increase in inflation and unemployment during 1973–5 and 1979–81, is to miss a crucial implication of modern Keynesian models with their lagged-inflation augmented Phillips wage equation, and raw-material price sensitive price equation. If the recently experienced rate of inflation is unacceptably high, or if the economy is subjected to a large upward supply-price shock (such as the dramatic increase in the price of oil in 1973–4 and again during 1979) then, modern Keynesian models assert, there will not exist *any* conventional fiscal and monetary actions that would produce cheerful results with respect to both (1) output and employment and (2) inflation. The entire slate of output–employment/inflation choices faced by the Federal Reserve, and Presidents Ford, Carter and Reagan was uninviting. Lacking an effective policy of direct price and wage restraint, Ford and Carter (and the Fed) could have achieved lower rates of inflation only at the cost of still more lost output and more (transient) unemployment. Reagan and Volcker could have achieved the President's ambitious 1981 output and employment objectives only at the cost of

persistently rapid inflation. (The NCM model's only explanation for the acceleration of inflation during the mid- and late 1970s is that the Federal Reserve became unhinged. A determined, well publicized policy of monetary restraint could have prevented any speed-up in inflation at virtually no cost in output and employment. That same model says that the Fed can near-costlessly stop inflation. Keynesian models assert that the cure is costly, as in fact it turned out to be during 1981–4.)

Remarks

Trade-offs involving inflation and unemployment will plague policymakers even in an accelerationist, natural rate, lagged-inflation augmented Phillips/Keynes world, especially one beset by upward supply-price shocks. The slate of inflation-unemployment choices in such a Phelps/Friedman/Phillips/Keynes economy is more complicated than in an old-fashioned Phillips/Keynes economy of the sort that Walter Heller had in mind in the early 1960s (perhaps correctly, for the range of $\dot{P}$ actually experienced during 1958–64 – there is no way to know). But only if prices instantaneously clear all markets, and, secondarily, if expectations are entirely free of inertia and strategic interdependence – that is if the economy is NCM in its structure – will the aggregate supply curve in $\dot{P}$-Q space be vertical in what may otherwise be a long-protracted short-run. (In NCM models, only capricious, unpredictable government actions give rise to an inflation-unemployment trade-off.)

One can espouse an actively responsive policy of demand management without condoning inflation. Preferences with respect to $\dot{P}$, $\ddot{P}$,..., Q and U bear on the choice of an aggregate demand target, not on how actively responsive the government should be in pursuing that target. There is no presumption that managers instructed to minimize inflation in a cost-effective manner would enjoy a quieter life than if they were told to favour output at the expense of faster inflation.

In a non-classical, Keynesian world, policy should aim at *both* nominal and real magnitudes, in a way that recognizes their interactions. An exclusively *nominal* strategy designed to yield a given year-to-year increase in nominal GNP (ΔPQ), no matter how it divides between increased prices (ΔP) and increased output (ΔQ) makes no sense whatever. The point is especially important if supply–price disturbances are important. *Real* targeting, if interpreted to mean that one should ignore inflation, is not acceptable either, unless one simply does not care about inflation *per se*, and about whatever microeconomic inefficiency it causes.

Theoretical considerations bearing on sensible portfolio behaviour, and evidence concerning the interest-responsiveness of the demand for money, make, I think, untenable the old monetarist claim that, even in the short run, *only* money matters – that fiscal action has no independent effect on total spending. With respect to the very long run, one has to be open-minded. The answer depends on the effect of the interest rate on the demand for wealth, i.e. on saving, and the effect of wealth on the demand for money. But that long run, equilibrium-to-equilibrium outcome seems to be of no practical significance.

The selection of a policy mix – from among the many combinations of budget settings and base-money growth compatible with one's preferred output and inflation target – should reflect the community's preferences with respect to the distribution of income and the division of output between consumption and investment, private and public. In other ways, too, policy should pay attention to supply as well as demand – how to get more output out of given capital and labour, and whether and how to upgrade and augment the former, and enhance the performance and pleasure of the latter.

Sensible managers will make tactical use of *any* intermediate indicator (e.g. free reserves, help wanted ads, Michigan surveys, whatever), as long as it exhibits sufficient short-run predictive power to improve their performance. But they will never waste degrees of freedom by treating such auxiliary aiming points as though they were objectives. They will avoid shibboleth goals like budget balance. Instruments are scarce enough, even relative to true objectives.

Because the American economy has become much more 'open', demand management in the US is more complicated than it was two decades

ago. The causal interconnections are more uncertain, and instruments are scarcer relative to targets. But that is not an argument for setting the controls on 'automatic'. Rather, it strengthens the case for an eclectic, regret-minimizing activism.

See Also

- ▶ Rational Expectations
- ▶ Targets and Instruments

Bibliography

Bator, F.M. 1982. Fiscal and monetary policy: In search of a doctrine. In *Economic choices: Studies in tax/fiscal policy*. Washington, DC: Center for National Policy.

Blinder, A.S., and R.M. Solow. 1984. Analytical foundations of fiscal policy. In *Economics of public finance*. Washington, DC: Brookings Institution.

Council of Economic Advisers. 1962. Annual report of the Council of Economic Advisers. *Economic Report of the President*. Washington, DC: US Government Printing Office.

Friedman, M. 1948. A monetary and fiscal framework for economic stability. *American Economic Review* 38: 245–264.

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

Heller, W.W. 1967. *New dimensions of political economy*. New York: Norton.

Lerner, A.P. 1941. The economic steering wheel. *University Review*, Kansas City, June, 2–8.

Lucas, R. 1976. Econometric policy evaluation: A critique. *Journal of Monetary Economics BTX Supplement, Carnegie-Rochester Conference Series* 1: 19–46.

Lucas, R. 1977. Understanding business cycles. *Journal of Monetary Economics, Supplement, Carnegie-Rochester Conference Series* 5: 7–29.

Lucas, R. 1980. Methods and problems in business cycle theory. *Journal of Money, Credit and Banking* 12(4, Pt II): 696–715.

Modigliani, F. 1977. The monetarist controversy, or, should we foresake stabilization policies? *American Economic Review* 67(2): 1–19.

Okun, A.M. 1971. Rules and roles for fiscal and monetary policy. In *Issues in fiscal and monetary policy: The eclectic economist views the controversy*, ed. James J. Diamond. Chicago: Depaul University Press. Reprinted in *Economics for policymaking, selected essays of Arthur M. Okun*, ed. Joseph Pechman. Cambridge, MA: MIT Press, 1983.

Okun, A.M. 1980. Rational-expectations-with misperceptions as a theory of the business cycle. *Journal of Money, Credit and Banking* 12(4, Pt II): 817–825.

Phelps, E.S. 1968. Money-wage dynamics and labor-market equilibrium. *Journal of Political Economy* 76(4, Pt II): 678–711.

Samuelson, P.A. 1951. Principles and rules of modern fiscal policy: A neo-classical reformulation. In *Money, trade and economic growth: Essays in honor of John Henry Williams*, ed. Hilda L. Waitzman. New York: Macmillan.

Samuelson, P.A., and R.M. Solow. 1960. Analytical aspects of anti-inflation policy. *American Economic Review* 50: 177–194.

Sargent, T.J., and N. Wallace. 1975. 'Rational' expectations, the optimal monetary instrument, and the optimal money supply rule. *Journal of Political Economy* 83(2): 241–254.

Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48(1): 1–48.

Solow, R.M. 1976. Down the Phillips curve with gun and camera. In *Inflation, trade and taxes*, ed. David A. Belsey et al. Columbus: Ohio State University Press.

Solow, R.M. 1979. Alternative approaches to macroeconomic theory: A partial view. *Canadian Journal of Economics* 12(3): 339–354.

Solow, R.M. 1980. What to do (macroeconomically) when OPEC comes? In *Rational expectations and economic policy*, ed. Stanley Fisher. Chicago: University of Chicago Press.

The Annual Report of the Council of Economic Advisers. 1962. *Economic Report of the President*. Washington, DC: US Government Printing Office.

Tobin, J. 1977. How dead is Keynes? *Economic Inquiry* 15(4): 459–468.

Tobin, J. 1980a. Are new classical models plausible enough to guide policy? *Journal of Money, Credit and Banking* 12(4, Pt II): 788–799.

Tobin, J. 1980b. Stabilization policy ten years after. *Brookings Papers on Economic Activity* 1(10th Anniversary Issue): 19–71.

Tobin, J. 1982. Steering the economy then and now. In *Economics in the public service*, ed. Joseph A. Pechman. New York: W.W. Norton.

Tobin, J. 1985. Theoretical issues in macroeconomics. In *Issues in contemporary macroeconomics and distribution*, ed. George Feiwel. New York: State University of New York.

Finite Sample Econometrics

Aman Ullah

Keywords

Asymptotic theory (large sample) econometrics; Bootstrap; Edgeworth expansion;

Edgeworth, F.; Empirical likelihood estimators; Finite sample method in econometrics; Generalized least squares estimators; Generalized method of moments estimators; Hypothesis testing; Least squares estimators; Likelihood ratio method; Linear models; Maximum likelihood estimators; Monte Carlo methods; Quantile estimators; Rao's score; Simultaneous equations models; Wald's test

JEL Classifications
C1

Economic models, which provide relationships between economic variables, are useful in making scientific predictions and policy evaluations. Well-known examples include classical linear regression models, where the explanatory variables are assumed to be non-stochastic (fixed) and the errors are normally distributed, and non-classical models, where these assumptions are violated. These non-classical models are frequently used in empirical work, and they include the simultaneous equations model, models with serial correlation and heteroscedasticity, limited dependent-variables models, panel and spatial models, non-linear models, and models with non-normal errors.

Based on sample data, econometric methods provide techniques of estimation and hypothesis testing related to these and other models. The commonly used estimators are the least squares (LS) or the generalized LS (GLS), the maximum likelihood (ML), the generalized method of moments (GMM), the empirical likelihood (EL) and the quantiles. The hypothesis-testing procedures used are Wald's (W), Rao's score (RS) and the likelihood ratio (LR) methods. Since all these are based on sample information, the statistical properties (unbiasedness, consistency, efficiency, distributions) of these procedures are of great interest for both small and large samples. This has led to the development of asymptotic theory (large sample) econometrics (White 2001) and finite sample econometrics (Ullah 2004).

The large sample theory properties may not imply finite sample behaviour of econometric estimators and test statistics, and they can give misleading results for small or even moderately large samples. As an example, consider a regression model

$$y_i = x_i\beta + u_i \qquad i = 1, 2..., n,$$

where y_i is a univariate response, x_i is a univariate fixed regressor, β is an unknown parameter to be estimated, and u_i is an additive error assumed to be independently and identically distributed (i.i.d.) with mean zero and variance σ^2. Let b_1, and $b_2 = (1 - 1{=}n)b_1$ be two estimators of β, where b_1 is the LS estimator. Then, the asymptotic distributions of b_1 and b_2 are

$$\sqrt{n}(b_1 - \beta) \sim N\left(0, \sigma^2/m_{xx}\right),$$
$$\sqrt{n}(b_2 - \beta) \sim N\left(0, \sigma^2/m_{xx}\right),$$

where $m_{xx} = \Sigma_{i=1}^{n} \frac{x_i^2}{n}$ as n tends to ∞.

Thus, asymptotically, both estimators are unbiased, and they have the same variances and distributions. But these results do not hold for finite samples (small or moderately large), since in this case $Eb_1 = \beta$, $Eb_2 = \beta(1 - 1/n)$, $V(b_1) = \sigma^2/\Sigma_{i=1}^{n} x_i^2$, $V(b_2) = (1 - 1/n)^2 V(b_1)$, that is, while b_1 is unbiased, b_2 is biased and their variances are different. Further, the distributions of b_1 and b_2 are generally not known but, if we assume normality of errors, then both b_1 and b_2 are normally distributed.

Fisher (1921, 1922) and then the work of Cramér (1946) laid the foundations of statistical finite sample theory on the exact distributions and moments which are valid for any sample size. This exact theory on distributions and moments was brought into econometrics by the seminal work of Haavelmo (1947) and Anderson and Rubin (1949) on the exact confidence regions of structural coefficients, Hurwicz (1950) on the exact LS bias in an autoregressive model, Basmann (1961) and Phillips (1983) on the exact density and moments of the estimators in the structural model, and Ullah (2004) on the exact moments. However, these exact results are often very complicated for drawing meaningful inferences since they are expressed in terms of multivariate integrals or complex infinite series. Also, the results are not derivable for non-classical models, especially for non-linear models or models with non-normal errors.

Another major development took place through the pioneering work of Nagar (1959) on obtaining the approximate moments of the k-class estimators in simultaneous equations. This was followed by Sargan (1975) and Phillips (1980), who rigorously developed the theory and applications of the Edgeworth expansions to derive the approximate distribution functions of econometric estimators. (The idea of the Edgeworth expansions originates from the fundamental work of Edgeworth 1896.) The approximate distributions and moments provide results which can tell us how much we lose by using asymptotic results and how far we are from the exact results if they are known. Most of the contributions, however, were confined to the analytical derivation of the moments and distributions in the simultaneous equations model and the dynamic first-order autoregressive (AR (1)) model, but with i.i.d. normal observations. These also included the finite sample results using the Monte Carlo methodology (Hendry 1984) and advances in bootstrapping (resampling) procedures (see Efron, 1979; Hall 1992). The analytical and bootstrap results for non-classical models, especially those that are non-linear with non-normal and non-i. observations, remain a challenging task for future development in this area of research. For the approximate analytical results some development has begun to take place (Rilstone et al. 1996) with a non-i.i.d. extension in Ullah (2004). This provides results which can be used to evaluate the approximate bias and mean-squared error of a class of estimators (ML, LS, GMM) for linear and non-linear models with normal or non-normal errors, and the observations can be i. i.d. or non-i.i.d. In the same spirit Newey and Smith (2004) develop the properties of generalized empirical likelihood estimators. Similarly, there are developments in the bootstrapping procedures for studying the properties of the GMM and extremum estimators in various econometric models with i.i.d. as well as dependent and non-stationary observations (see Horowitz 2001).

The progress in finite sample econometrics has indeed been ongoing. The developments described provide analytical and simulation-based procedures for finite sample analysis of econometric models. In the broad sense, the frontier of this research area has moved on. With the advances in computer technology this subject will further develop in both the analytical and the bootstrapping domains.

See Also

- ▶ Bootstrap
- ▶ Econometrics
- ▶ Simultaneous Equations Models

Bibliography

Anderson, T., and H. Rubin. 1949. Estimation of the parameters of a single equation in a complete system of stochastic equation. *Annals of Mathematical Statistics* 20: 46–63.

Basmann, R. 1961. Note on the exact finite sample frequency functions of generalized classical linear estimators in two leading overidentified cases. *Journal of the American Statistical Association* 56: 619–636.

Cramér, H. 1946. *Mathematical methods of statistics*. Princeton: Princeton University Press.

Edgeworth, F. 1896. The asymmetrical probability curve. *Philosophical Magazine* 41: 90–99.

Efron, B. 1979. Bootstrap methods: Another look at the jacknife. *Annals of Statistics* 7: 1–26.

Fisher, R. 1921. On the probable error of a coefficient of correlation deduced from a small sample. *Metron* 1: 1–32.

Fisher, R. 1922. The goodness of fit of regression formulae and the distribution of regression coefficients. *Journal of the Royal Statistical Society* 85: 597–612.

Hall, P. 1992. *The bootstrap and edgeworth expansion*. New York: Springer-Verlag.

Haavelmo, T. 1947. Methods of measuring the marginal propensity to consume. *Journal of the American Statistical Association* 42: 105–122.

Hendry, D. 1984. The Monte Carlo experimentation in econometrics. In *Handbook of econometrics*, ed. M. Intriligator and Z. Griliches, Vol. 2. Amsterdam: North-Holland.

Horowitz, J. 2001. The bootstrap in econometrics. In *Handbook of econometrics*, ed. J. Heckman and E. Leamer, Vol. 5. Amsterdam: North-Holland.

Hurwicz, L. 1950. Least square bias in time series. In *Statistical inference in dynamic economic models*, ed. T. Koopmans. New York: Wiley.

Nagar, A. 1959. The bias and moments matrix of the general k-class estimators of the parameters in structural equations. *Econometrica* 27: 575–595.

Newey, W., and R. Smith. 2004. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* 72: 219–255.

Phillips, P. 1980. Finite sample theory and the distribution of alternative estimators of the marginal propensity to consume. *Review of Economic Studies* 47: 183–224.
Phillips, P. 1983. Exact small sample theory in simultaneous equations models. In *Handbook of econometrics*, ed. M. Intriligator and Z. Griliches, Vol. 1. Amsterdam: North-Holland.
Rilstone, P., V. Srivatsava, and A. Ullah. 1996. The second order bias and MSE of nonlinear estimators. *Journal of Econometrics* 75: 239–395.
Sargan, J. 1975. Gram-Charlier approximations applied to t ratios of k-class estimators. *Econometrica* 43: 326–346.
Ullah, A. 2004. *Finite sample econometrics*. New York: Oxford University Press.
White, H. 2001. *Asymptotic theory for econometricians*. New York: Academic Press.

Finley, Moses (1912–1986)

Isabel Raphael

Keywords
Finley, M.; Slavery

JEL Classifications
B31

Sir Moses Finley had an immense influence on classical studies and particularly ancient history because he brought to them the new disciplines and techniques of the modern social sciences. He was unique among ancient historians in that his early training had been in law, economics and sociology.

Born on 20 May 1912, Finley graduated (BA) from Syracuse University at the age of 15 and from Columbia (MA) at 17, his major subjects being psychology and US constitutional law. Westermann encouraged him to try ancient history, and he taught himself Latin and Greek, financing himself with his earnings and those of his wife Mary, a school teacher whom he married in 1932. Theirs was a childless but devoted marriage, Lady Finley dying two days before her husband.

Finley worked from 1930 to 1933 on the *Encyclopedia of Social Sciences* and was much influenced by the Frankfurt Institute for Social Research; his reading of social theory made him left-wing and at least partly Marxist. He was active on behalf of the Republicans during the Spanish civil war and raised funds for Russian war relief in the Second World War. After founding the American Committee for the Defence of International Freedom against McCarthyism he was dismissed from his post as Assistant Professor of History at Rutgers University. Known by now for his lectures in England, he was given the post of lecturer in classics at Cambridge in 1955, and was a Fellow of Jesus College from 1957 to 1976. He became a British subject in 1962. He succeeded to the chair of ancient history in 1960, and in 1976 became the first Master of Darwin College. Finley's doctoral dissertation, 'Studies in Land and Credit in Ancient Athens' (1950), gained him an international reputation. He asked questions that had not been considered before in this field, and saw the ancient world with modern eyes. Classical scholars had used the word 'economics' in its ancient and particular sense, as the management of a household and hence of a state; Finley opened up the discipline to the interests of modern social sciences, dealing with matters such as property, contracts, succession, the value of goods and coin and the laws of war. He stepped aside from the traditional track to look at the exact relationship between masters and slaves, the nature of debt bondage, the consumer society and urban and rural production. He was the first ancient historian to tackle the methodological problems implied by the new style of social history.

Finley could appear cantankerous and was famous for his feuds; he enjoyed creating shock waves in the academic world. But at his best he was a new wind blowing through an old and rather old-fashioned subject, and he changed and refreshed the classics more than any other scholar this century.

Selected Works

1956. *The world of Odysseus.* London: Chatto & Windus.

1963. *The ancient Greeks.* London: Chatto & Windus.
1970. *Early Greece: The bronze and archaic ages.* London: Chatto & Windus.
1973a. *Democracy ancient and modern.* London: Chatto & Windus. Revised ed, 1985.
1973b. *The ancient economy.* London: Chatto & Windus.
1980. *Ancient slavery and modern ideology.* London: Chatto & Windus.

Firm Boundaries (Empirical Studies)

Thomas N. Hubbard

Abstract

The empirical literature on the determinants of firms' boundaries examines relationships between firms' boundaries and asset specificity, especially how relationship-specific investments create 'hold-up' problems that increase the costs of competitive contracting; relationships between firms' boundaries and the contracting environment, reflecting the role of incomplete contracting in the theoretical literature and the extent to which firms subcontract downstream stages rather than input procurement; and how firms' boundaries vary with 'job design'. This literature has established that asset specificity is empirically relevant for understanding integration decisions, and that relationships between subcontracting decisions, the contracting environment, and the division of labour are subtle.

Keywords

Agency costs; Asset specificity; Coase, R.; Contracting; Division of labour; Firm boundaries; Hold-up problem; Incomplete contracts; Outsourcing; Vertical integration

JEL Classifications

L22

This article discusses empirical work on the determinants of firms' boundaries, focusing on 'make-or-buy' decisions. Examples of such decisions include whether firms procure inputs (or distribute outputs) through in-house divisions or other firms. It concentrates on work that draws on Coase (1937), which depicts firms and markets as alternative means of governing transactions. The theoretical literature in the Coasean tradition is vast, and includes well-known works by Williamson (1975, 1979, 1985), Klein, Crawford and Alchian (1978), Grossman and Hart (1986), and Holmstrom and Milgrom (1994). This contrasts with the neoclassical literature, in which firms' boundaries are determined by production technology and, perhaps, market power-related issues. This other literature examines how, for example, vertical integration reflects firms' incentive to eliminate double marginalization, raise rivals' costs, or protect themselves from competitors' attempts to raise their own costs.

Firms' Boundaries and Relationship-Specific Assets or Investments

By far the largest branch of the empirical literature examines relationships between firms' boundaries and asset specificity. This branch is primarily motivated by Klein, Crawford and Alchian's (1978) and Williamson's (1979, 1985) analysis of how relationship-specific investments create 'hold-up' problems that increase the costs of competitive contracting. On the assumption that such investments do not create as severe problems when transactions take place within firms, it follows that vertical integration should be more prevalent, and outsourcing less prevalent, when transactions involve relationship-specific assets than when they do not.

Several early papers examine this proposition in procurement contexts. Monteverde and Teece (1982) and Masten (1984) examine outsourcing decisions of auto makers and an aerospace firm, respectively, and find that outsourcing is less prevalent when components are firm-specific than not. The latter finds that it is also less prevalent when

co-locating production of the component with that of successive production stages is more valuable. Joskow (1985) finds that vertical integration is prevalent when coal-burning power plants are located close to coal mines, but power plants procure coal from outside firms when they are not co-located. These and other correlations uncovered by this early work provided the first evidence that asset specificity is empirically relevant for understanding integration decisions and, more broadly, that analysing firms' boundaries from a contractual perspective could lead to new empirical insights.

This branch has since developed along several lines. Researchers have found relationships between asset specificity and vertical integration in other industrial contexts, and have explored the empirical limits of this proposition by examining the extent to which asset specificity and integration are correlated in contexts where investments are smaller and less specific than in the contexts discussed above. Still others have investigated the closely related question of how, given that vertical integration is not chosen, contractual relationships vary with asset specificity (see Joskow 1988, and Klein 2005, for comprehensive surveys).

There is significant debate over the theoretical interpretation of this evidence. Asset specificity is an important element of many theories in this literature, so correlations between asset specificity and vertical integration need not provide evidence in favour of any one in particular. Whinston (2003) discusses this problem at length, and concludes that, while these theories' empirical implications are not the same, the data requirements of distinguishing tests are considerable and the existing empirical evidence is not dispositive.

Some recent papers indicate that the relationship between vertical integration and investment can be subtle. Woodruff's analysis (2002) of vertical integration between shoe manufacturers and retailers, and Acemoglu et al.'s analysis (2004) of vertical integration in British manufacturing indicate that whether vertical integration is more or less prevalent when investments are more important depends critically on the source and nature of the investment. Understanding empirical relationships between integration and investment incentives is a major focus of current research.

Firms' Boundaries and the Contracting Environment

A second branch of the empirical literature examines relationships between firms' boundaries and the contracting environment. Many theories motivate this research, reflecting the essential role incomplete contracting plays throughout the theoretical literature. This branch typically examines the extent to which firms subcontract downstream stages rather than input procurement. Examples include Anderson and Schmittlein's work (1984) on whether manufacturers rely on internal or external sales representatives, Baker and Hubbard's investigations (2003, 2004) of firms' boundaries in trucking, Brickley, Linck and Smith's analysis (2003) of whether bank offices are independent entities or branches, and some of the research (see Lafontaine and Slade 1997, for a survey) that examines whether chain outlets are company- owned or franchises. These papers exploit variation in the availability of good measures of downstream individuals' performance, which, in turn, derives from technological change or differences in the nature of the downstream individual's job. (Work exploiting the latter is classified here rather than below, when authors emphasize differences in the contractibility rather than the number or diversity of tasks.) Results from these papers generally indicate that more (downstream) integration tends to be associated with better performance measures.

These results have several implications. First, they imply that the contracting environment is not organization-neutral. Non-neutrality is not obvious. Agency problems exist between upstream and downstream entities regardless of whether the latter are employees or subcontractors. If contractual improvements reduce agency costs independently of integration-related trade-offs, they should not affect firms' boundaries. The results indicate otherwise. Second, they imply that the contracting environment affects the costs of transacting within as well as between firms. Some

theories, including Coase (1937), propose that coordination takes place 'by fiat', and hence the contracting environment is irrelevant, within firms. If so, contractual improvements should always favour market transacting and thus less vertical integration. Again, the results indicate otherwise. Third, they suggest that while contracting problems exist both within and between firms, empirical variation in the availability of good performance measures tends to be related to inefficiencies associated with transacting within firms. Although this conclusion is preliminary, it implies that it is particularly productive for those researching (or making) 'make-or- buy' decisions to identify the source of these inefficiencies, because they may have more real-world 'bite'.

Firms' Boundaries and the Division of Labour

A third, related branch examines how firms' boundaries vary with the division of labour, or 'job design'. Holmstrom and Milgrom's (1994) analysis of how multitask agency problems influence firms' boundaries motivates much of this work. This branch includes analyses of how whether in-house salesmen or sales reps are used depends on whether salesmen are also given non-selling responsibilities (Anderson 1985), how whether pharmaceutical firms outsource clinical trials depends on whether the work involves more than just data collection (Azoulay 2004), and how whether restaurants are company-owned or franchised depends on how much food production and service takes place at the restaurant (Yeap 2005).

Results from these papers indicate that integration tends to be less prevalent when individuals are allocated a narrower set of responsibilities. Combined with the evidence above, they imply that relationships between subcontracting decisions, the contracting environment, and the division of labour are subtle. The previous subsection suggests that *replacing* an easily contractible task with a less contractible one tends to make subcontracting more likely. The evidence here suggests that *adding* a less contractible task to a more contractible one tends to make subcontracting less likely.

Other work has found that the division of labour and firms' boundaries are related in horizontal contexts as well; for example, Garicano and Hubbard (2003) find that law firms' field boundaries narrow as market size increases and lawyers become more specialized.

Firms' Boundaries and Economic Outcomes

Most of the literature investigates what determines whether firms integrate rather than what actually happens when they do, but research on the latter is important because it reveals whether integration is an economically important issue.

Some evidence has come from firm or industry case studies. Early work includes Masten, Meehan and Snyder (1991), which concludes that organizational costs make up a significant fraction of production costs in shipbuilding, and that incorrect choices with respect to integration decisions can increase organization- related costs by as much as 70 per cent. More recently, Gil (2004) investigates relationships between how long movies play at a theatre and whether the theatre is owned by the movie's distributor, and finds that movies play two weeks longer at distributor-owned theatres than other, similarly situated theatres.

Perez-Gonzalez (2004) provides cross-industry evidence. This investigates how the elimination of Mexican laws that constrained multinational firms from having majority control of affiliated enterprises affected plant-level investment and productivity. He finds that, within technology-intensive sectors, increases in integration associated with the elimination of these constraints led to significant investment increases and an approximately ten per cent increase in total factor productivity at these enterprises. The allocation of control rights, and thus vertical integration, can have a major impact on investment incentives and productivity. In short, integration decisions can matter a lot.

See Also

- ▶ Contract Theory
- ▶ Hold-up Problem
- ▶ Incomplete Contracts
- ▶ Property Rights

Bibliography

Acemoglu, D., P. Aghion, R. Griffith, and F. Zilibotti. 2004. *Vertical integration and technology: Theory and evidence*, Working paper No. 10997. Cambridge, MA: NBER.

Anderson, E. 1985. The salesperson as outside agent or employee: A transaction-cost analysis. *Marketing Science* 4: 234–254.

Anderson, E., and D. Schmittlein. 1984. Integration of the sales force: An empirical examination. *Rand Journal of Economics* 15: 385–395.

Azoulay, P. 2005. Capturing knowledge within and across firm boundaries: Evidence from clinical development. *American Economic Review* 94: 1591–1612.

Baker, G., and T. Hubbard. 2003. Make versus buy in trucking: Asset ownership, job design, and information. *American Economic Review* 93: 551–572.

Baker, G., and T. Hubbard. 2004. Contractibility and ownership: On-board computers and governance in U.S. trucking. *Quarterly Journal of Economics* 119: 1443–1479.

Brickley, J., J. Linck, and C. Smith. 2003. Boundaries of the firm: Evidence from the banking industry. *Journal of Financial Economics* 70: 351–383.

Coase, R. 1937. The nature of the firm. *Economica* 4: 386–405.

Garicano, L., and T. Hubbard. 2003. *Specialization, firms, and markets: The division of labor within and between law firms*, Working paper No. 9719. Cambridge, MA: NBER.

Gil, R. 2004. *Decision rights and vertical integration in the movie industry.* Mimeo. Department of Economics, University of California, Santa Cruz.

Grossman, S., and O. Hart. 1986. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy* 94: 691–719.

Holmstrom, B., and P. Milgrom. 1994. The firm as an incentive system. *American Economic Review* 84: 972–991.

Joskow, P. 1985. Vertical integration and long term contracts: The case of coal- burning electric generating plants. *Journal of Law, Economics, and Organization* 1: 33–80.

Joskow, P. 1988. Asset specificity and the structure of vertical relationships: Empirical evidence. *Journal of Law, Economics, and Organization* 4: 95–117.

Klein, P. 2005. The make-or-buy decision: Lessons from empirical studies. In *Handbook of new institutional economics*, ed. C. Menard and M. Shirley. Boston: Kluwer.

Klein, B., R. Crawford, and A. Alchian. 1978. Vertical integration, appropriable rents and the competitive contracting process. *Journal of Law and Economics* 21: 297–326.

Lafontaine, F., and M. Slade. 1997. Retail contracting: Theory and practice. *Journal of Industrial Economics* 45: 1–25.

Masten, S. 1984. The organization of production: Evidence from the aerospace industry. *Journal of Law and Economics* 27: 403–417.

Masten, S., J. Meehan, and E. Snyder. 1991. The costs of organization. *Journal of Law, Economics, and Organization* 7: 1–25.

Monteverde, K., and D. Teece. 1982. Supplier switching costs and vertical integration in the automobile industry. *Bell Journal of Economics* 13: 206–213.

Perez-Gonzalez, F. 2004. The impact of acquiring control on productivity. AFA 2005 Philadelphia Meetings.

Whinston, M. 2003. On the transaction cost determinants of vertical integration. *Journal of Law, Economics, and Organization* 19: 1–23.

Williamson, O. 1975. *Markets and hierarchies: Analysis and anti-trust implications*. Glencoe: Free Press.

Williamson, O. 1979. Transaction-cost economics: The governance of contractual relations. *Journal of Law and Economics* 22: 233–261.

Williamson, O. 1985. *The economic institutions of capitalism*. New York: Free Press.

Woodruff, C. 2002. Non-contractible investments and vertical integration in the Mexican footwear industry. *International Journal of Industrial Organization* 20: 1197–1224.

Yeap, C. 2005. *Residual claims and incentives in restaurant chains*. Mimeo. University of Chicago.

Firm, Theory of the

G. C. Archibald

Keywords

Agency theory; Bargaining; Bounded rationality; Comparative statics; Cournot, A. A.; Creative destruction; Dynamic programming; Expense preference; Firm, theory of the; Free-rider problem; Gibrat's Law; Implicit contracts; Incentive compatibility; Innovation; Institutional rent; Insurance; Joint-stock companies; Laws of returns; Linear programming;

Long run and short run; Long-run equilibrium; Marginal revolution; Markov processes; Monitoring; Optimal control; Optimizing models; Ownership and control; Principal and agent; Quasi-rent; Rent; Representative firm; Residual; Returns to scale; Risk; Risk aversion; Risk sharing; Transaction costs; Uncertainty

JEL Classifications
D2

It is doubtful if there is yet general agreement among economists on the subject matter designated by the title 'theory of the firm', on, that is, the scope and purpose of the part of economics so titled. There is, probably, general agreement on the subject matter of economics itself: the allocation and distribution of scarce resources. (Some economists would have us add explicitly 'and growth' to 'allocation and distribution', but traditionally growth is subsumed under 'allocation'.) Then we may take it that the purpose of the theory of the firm is to investigate the behaviour of firms as it affects allocation and distribution. We now come immediately to a fork. An economist who believes that a 'firm' is a profit-maximizing agent (whether by conscious, rational decision or otherwise), endowed with a known and given technology, and operating subject to a well-defined market constraint, will see no need for any special theory of the firm: the theory of the firm is nothing but the file of optimizing methods (and perhaps market structures). *If* firms maximize, *how* they do it is not of great interest or at least relevance to economics. The economist's job is simply to cultivate and apply optimizing techniques. Given this view, it is unnecessary to inquire further: to seek to 'inquire within' is otiose, perhaps methodologically misguided. (As we shall see, the theory of the firm has been, and perhaps still is, the battleground for some fierce methodological warfare.)

Economists who doubt any of the three critical assumptions see an urgent need to inquire within, but diverge substantially thereafter (for example, managerial utility functions, behaviourism). Later on, I shall try to exhibit a systematic tree, although this is not easy since some of the branches are sadly tangled. Before doing that, I want to show that the first fork, referred to above, was recognized a long time ago, and to sketch some of the history of our subject. First, though, I must impose more narrow limits on it.

In most of the work on the theory of the firm it is at least implicitly assumed that the agent whose behaviour is to be examined is a capitalist firm (which may or may not be a joint-stock corporation) engaged in manufacturing, processing or perhaps extraction. Thus the study of financial intermediaries, although they are firms, is conventionally relegated to some other branch of our discipline. Partnerships and cooperatives (labour-managed firms) may be usefully examined with the techniques of the theory of the firm, as may not-for-profit organizations, but their study is conventionally filed under 'comparative systems'. For convenience and brevity, although not out of conviction, I shall respect these conventions here. It is also necessary to place some demarcation line between the theory of the firm and 'market structure' or 'industrial organization'. For the moment, at least, I think it better to let this one be implicit.

We must also ask why firms exist at all. The classic – and neoclassical – answer was provided by Coase (1937): transactions costs. I call this a 'neoclassical' answer because part of the tradition, still embodied in much contemporary general equilibrium theory, is the assumption of constant returns to scale. Some increasingness of returns may be a very good reason for the existence of firms, or at least help to explain their size, but it is obviously vastly convenient to have a sufficient reason which is not inconsistent with constant returns. Coase suggested that the firm was an area (subset of the economy) in which allocation proceeded by direction rather than via markets, because some procedures, such as the allocation of workers to tasks, could be more cheaply done that way – coordination by command rather than by price. The word 'command' suggests that some monitoring, enforcement or internal incentive structure will be required, and indeed these matters have been receiving increasing attention. Alchian and Demsetz (1972), in particular, discussed the problem of monitoring, suggesting,

in effect, that the need for it explained and justified the existence of the capitalist firm. They posed the question of who monitors the monitor, and suggested that the incentive problem is solved if the ultimate monitor is the residual claimant. O. Williamson (1980) reviewed alternative organizational structures. He suggested that the existence of firms economizes on explicit contracts which, given uncertainty and bounded rationality, are expensive instruments. He also found that ownership and hierarchy are only weakly related.

A recent work to emphasize the reasons for the existence of firms is Aoki's (1984). He argues that if firms exist because institutional allocation is cheaper than market allocation, reasons for which he explores thoroughly, then firms must enjoy 'institutional rent'. Furthermore, not all the resources used within the firm will have prices uniquely determined by external markets. Thus the distribution of rewards is not uniquely determined, and there is room for bargaining. Aoki argues that this is best modelled as a cooperative *game,* the players of which are the stockholders and the workers. Managers are reduced to the role of technocratic mediators (which, in view of recent developments in agency theory, discussed below, is perhaps surprising). This approach proves to be very flexible: Aoki can handle as special cases the neoclassical model (shareholders get all the residual) and the labour-managed firm in which the workers get it all (and even, with some interpretation, managerial models).

In what follows, I shall take the existence of firms for granted and return later to the matter of incentives.

The first fork, referred to above, will be familiar to any careful reader of Adam Smith (1776). He relied upon the self-interest of the butcher, the baker and the brewer to provide his dinner. The 'firms' in which he had confidence were small, owner-operated (whether single owner or partnership), without limited liability. He had serious misgivings about joint-stock companies. He pointed out what has become known in this century, thanks to Berle and Means (1933), as the 'divorce between ownership and control'. And he doubted if the managers had appropriate incentives to try to maximize the owners' returns; that is, he raised the question of what is now called 'incentive compatibility'. Thus, in considering the joint-stock company, Smith went unhesitatingly down what I will call the 'troublemaker's branch': we do have to inquire within. The joint-stock company is, of course, the predominant contemporary organization.

After Smith, there is not much that can be called 'theory of the firm' in classical economics. (Ricardo's firms are Smith's butchers and bakers.) The exception, as so often, is Marx, but there is not space to discuss Marx here. (J.S. Mill 1848, in the famous chapter 'On the Probable Futurity of the Labouring Classes', expressed concern about both the incentive structure and morality of the capitalist form of organization, and recommended a cooperative form instead.) We must notice, however, the startlingly modern work of Cournot (1838). He wrote down a demand function and, in his famous discussion of the mineral spring, employed explicit optimizing methods (and, so far as I know, was the first to do so). Not only this, he carried out a deliberate and formal exercise in comparative statics – in 1838! In applying marginal analysis to the theory of the firm he thus thoroughly anticipated the 'marginalists'. The 'marginal revolution' in due course produced a wholly desirable unification of the theories of production, allocation and distribution, creating the neoclassical branch from the fork, but with little that could be called 'theory of the firm'. The firm was, however, central in Marshall's (1890) work, and he, characteristically, put a foot on each branch. Formal, mathematical, Marshall is strictly neoclassical, as I employ the term. The informal Marshall, concerned with growth, offered suggestive literary dynamics.

Let us consider first the more formal Marshall. His distinction between the short and long runs is essential to much of his work. This distinction is, of course, the one currently in use: in the long run all factors are variable, in the short run one at least (commonly capital) is not. This allowed him to distinguish between fixed and variable costs, and between the effects of adding more labour to a fixed-capital stock and the effects of altering the scale of operations. We now have short-run diminishing returns in industry generally, while

there may be increasingness in the long run. Thus Marshall was not limited to the constant coefficients case of his classical predecessors: he was able to offer a thorough analysis of the 'laws' of returns. This allowed him to give a fairly complete analysis of the short-run equilibrium conditions for a firm selling in a perfect market. (There is in his analysis an even shorter 'short run', the market period in which the price of, say, a catch of herrings is determined. This does not appear to concern us here.) Marshall did not, of course, solve all the problems of the theory of production, costs, supply and distribution in competition. He left room for the important work of Viner (1931) and Stigler (1939).

A further and vital step was Marshall's generalization of Ricardo's theory of rent. He distinguished between a quasi-rent, which would in the long run be competed away, and a true rent, which definitionally could not be. (Both, of course, are any excess of rewards over opportunity cost.) If the quasi-rent is due to an increase in the demand for the product of specific capital equipment, then the long run in which it is competed away and the long run in which all factors are variable are, of course, identical. (That the period in which quasi-rent is competed away and that in which all factors are variable may differ is noted below.) This in turn allowed Marshall to develop the long-run equilibrium conditions for a competitive industry: quasi-rent must be competed away (or negative profit eliminated by exit) so that the normal profit condition is satisfied. Here he seems to have followed Walras (1874).

Marshall made many other contributions to the theory of the firm. He noted that, if increasingness in returns (to scale, as we should say) is internal to the firm, competition is not viable, whence a downward-sloping competitive supply curve can only be attributed to economies external to the firm (internal to the industry; but he also considered economies external to the industry and internal, perhaps, only to the whole economy). He also offered a formal monopoly model some features of which require remark. The firm's demand curve coincides with the market demand curve for the 'product' (a given primitive of analysis): there is no oligopolistic interaction here. This model is still with us, although the analysis has become more elegant. In his geometry, Marshall had us finding the profit-maximizing output by looking for the biggest profit rectangle: $(AR\text{-}AC)q$. Cournot (1838) had written down the marginal revenue function in his discussion of the mineral springs case, but Marshall chose not to follow him. (The discovery of the marginal revenue curve in Cambridge in the 1930s seems to have caused great excitement.)

The less formal Marshall was concerned with growth and the intertemporal behaviour of firms. His firms were joint-stock, but otherwise rather Smithian. He had, loosely speaking, a 'clogs to clogs in three generations' model. The first entrepreneur would be vigorous and innovative, finding some source of quasi-rent. His son would be more passive and probably mistake the quasi-rent for rent itself. The spoiled and idle grandson would certainly make this mistake, the quasi-rent would be competed away, and the cycle would be over.

This is, of course, not a good description of the history of a typical (immortal) joint-stock company. What is important is the link between innovation, quasi-rent and economic growth. Now, of course, the period in which quasi-rent is competed away is not necessarily identical to that in which capital can be varied. It may be possible to copy an innovation very quickly, or necessary to wait for the expiry of a patent. And if the quasi-rent is due to exceptional managerial talent and vigour (really, a rent to ability), it does not get competed away at all, but eventually withers. It was, however, this link between innovation and quasi-rent that Schumpeter (1934) made explicit in his great vision of the source of growth in a capitalist economy: the incessant seeking for quasi-rent via innovation, each source of quasi-rent being in turn competed away by further innovation in the process of 'creative destruction'. One notes, of course, that this model does not depend on the generational cycle of Marshall's family firm: widely owned joint-stock companies can continue to play Schumpeter's game so long as they are appropriately managed.

Marshall had the task of reconciling his view of the intertemporal behaviour of firms with his short-run profit-maximizing conditions and

long-run industry equilibrium conditions. His device of the 'representative firm' appears to have been designed to do this. The representative firm would not only be in short-run profit-maximizing equilibrium but would be earning precisely normal profit when the industry as a whole was in equilibrium. This means that the definition of long-run equilibrium needs to be more carefully stated. It is not 'all firms earn normal profit'. It is rather 'there is no tendency for the total number of firms in the industry to alter; the representative firm earns normal profits but others may still be expanding or already withering; in any case the net change is zero.' Here the representative firm is implicitly defined. As Newman (1960, p. 590) put it, in his discussion of Marshall's 'statistical' concept of long-run equilibrium, 'Long-run equilibrium for Marshall meant the equality of long-run demand and supply; just that and no more.' In the 1920s and 1930s there was a considerable literature on Marshall's value theory, not discussed here (see Newman 1960, for references). Since the work of Chamberlin (1933) and Joan Robinson (1933), the notion of the representative firm has tended to disappear from the literature. It has become usual to assume that each firm is always, by choice, in short-run equilibrium, and then to consider how Marshall's long-run competitive forces will impose industry equilibrium (normal profit for all firms simultaneously). Newman and Wolfe (1961), on the other hand, followed up the 'statistical' interpretation of Marshall's long-run equilibrium. They were not the first to apply Markov-chain analysis to the behaviour of an industry; but they were the first to integrate it with value theory. (Other more or less contemporary applications of Markov-chain analysis at most appeal to 'Gibrat's Law'. Newman and Wolfe may be thought to have prepared the ground for Nelson and Winter 1982, discussed below.)

I shall now attempt to describe some other forks and branches of the tree. To do this it is easiest to jump to the present, since so much has happened since the Second World War that needs to be allocated to its appropriate branch. (Chamberlin 1933, and Joan Robinson 1933, had, of course, made significant extensions of Marshall's formal models before the war. These contributions are discussed elsewhere.)

We encountered above a fork between what I call the smooth neoclassical branch and the rough and troublesome 'other' branch. There is another possible basis for classification, between optimizing and other models. The advantage of the first is that it gives the neoclassical model the prominence it deserves; the advantage of the second that it brings into prominence the importance of the assumptions we make about information and computational capacity. Perhaps somewhat arbitrarily, I shall classify the models to be considered here as optimizing and 'other'. The optimizing set of models divides again, between profit maximization and the optimization of other (usually managerial) objective functions.

Let us consider some arguments concerning the classes of models we have already identified.

The advantages of an optimizing model are clear: it is analytically tractable. We have well-developed techniques to handle it, even if the economic agents considered may not. It may also be thought to have important predictive power, but this is more dubious. The programme of qualitative comparative statics (Samuelson 1947) has been shown to be more limited than we might have hoped. The objections to optimizing models are well known, but also debatable. They are essentially two. The first is that firms, or the human beings that manage them, cannot optimize: they have neither the information nor the computational capacity, whence the most we can have is Simon's 'bounded rationality' (Simon 1955, 1959; see also 1979). Nelson and Winter (1982) have recently made a major contribution to this approach, discussed below. The position here is not that we give up the fundamental Smithian assumption of purposeful, self-interested behaviour (with what would we replace it?) but rather that we abandon the optimizing model and consider instead how, in a world of uncertainty, firms (managers) may explore their environment and try to 'make the best of it'. It is not suggested, at least by Nelson and Winter, that we 'inquire within' for the sake of it but rather to improve our understanding of how actual firms, seeking for profit but essentially too ignorant to optimize, may try to allocate resources. The second objection to optimizing models comes from those who have

enquired within and report that firms 'just don't' (see, for example, Hall and Hitch 1939; Andrews 1949; Cyert and March 1963). Many critics of this behaviourist school feel that it says little more than 'firms do what they do', and fails to analyse the relationship between the observed behaviour reported and resource allocation.

An example may show the force of the criticism. It is no longer open to doubt that firms commonly adopt markup pricing routines. In their study of a department store, Cyert and March (1963) report their discovery of the markup formula in use. They then congratulate themselves on being able to predict, given the wholesale price of an article, its posted price. They also notice that if profits are not satisfactory, the firm may adjust by altering its product-mix; that is, buying better (more expensive) or cheaper stock. But it is here that the important allocational decisions are taken, and this decision process is not analysed at all. (It should be noted that Cyert and March 1963, p. 268, place on their agenda matters which do not appear to be relevant to allocation and distribution at all, and which I accordingly exclude from consideration.)

Two related arguments in favour of profit-maximizing models may usefully be noticed now. The first is the 'biological analogy': survival of the fittest (see Alchian 1950; Penrose 1952; Friedman 1953; Machlup 1946, 1967). It is suggested that in a competitive world a firm must maximize to survive. Thus, however decisions are taken, whatever routines are adopted, firms which in fact maximize will prosper and be able, in particular, to retain and attract capital, while those that do not will wither. There are three points to raise here. The first is: how competitive is the environment? (See below.)

The second is that to survive, one does not have to be perfect but only good enough to handle the competition. Indeed, Charles Darwin seems to have anticipated this misuse of his argument when he wrote,

> Natural selection tends only to make each organic being as perfect as, or slightly more perfect than, the other inhabitants of the same country with which it has to struggle for existence ... Natural selection will not produce absolute perfection ... (Darwin 1859, pp. 201–2)

The third is that, to make effective use of the *biological* analogy, one has to offer something that can serve as a *gene.* Nelson and Winter (1982) have recently suggested a candidate (see below).

The second, and related, argument is that one can maximize without consciously trying. Thus Day and Tinney (1968) show that a firm can climb to the top of a (suitably concave) profit 'hill' by use of a simple feedback algorithm: if an action (change in output) succeeds (increases profit), repeat it; if not, back up. The notion that one may climb the hill 'driving only by the rear-vision mirror' must certainly be attractive to those who worry about the firm's information state and computational capacity. Yet obviously this simple feedback process works only if it converges ' fast enough' relative to the stability of the environment. Otherwise, it will be necessary to improve the algorithm to speed up convergence; for example, by adding feed-forward loops. The survival argument suggests that it will then be the firms that can do this that will survive. Then the loops (routines) are identified by Nelson and Winter as the genes in the evolutionary process. Notice, however, that this identification was made in 1982, not by those who originally proposed the biological analogy (see also Winter 1975).

We have now distinguished between optimizing models and 'other'. We have glimpsed the next two subdivisions, that between profit-maximizing and other optimizing models, and between behaviourism and other non-optimizing models. (We shall soon find another fork on the profit-maximizing branch, too; see below.) We have also noticed some relevant argument. We may now explore some developments along each of these branches.

Developments in and since the Second World War, some emerging from operations research, have extended the scope of optimizing models at a staggering rate. In a few short years, we had linear programming (for economic applications, see Dorfman et al. 1958), and activity analysis (see Koopmans 1951). Optimizing techniques were extended to inventory control (Whitin 1953; Simon 1952). We then had what I will call the 'dynamic explosion' as the techniques of

optimal control and dynamic programming were increasingly applied to the firm's problems; see, for example, Lucas (1967) and Treadway (1969) on the flexible accelerator, Mortensen (1970) and Brechling (1975) on the demand for labour.

Another major development has been the extension of optimizing models of the firm to include considerations of risk. Risk had been explicitly considered by Knight (1921), who offered an unsurpassed account of the ways in which the institutions of the capital market facilitate risk sharing. Knight tried to distinguish between 'risk' and 'uncertainty' in a way that many have found unsatisfactory: 'risk' was insurable; 'uncertainty', any uninsurable residual. Profit was the reward for bearing uncertainty (since risk could be covered by insurance). He was, I believe, the first to make the point that entrepreneurs would have to be less risk-averse than others (their employees) with whom they entered into explicit contracts. Recent work does not, however, follow Knight. It took a new departure from the work of von Neumann and Morgenstern (1944); see particularly Arrow (1971), and for specific applications to the theory of the firm, see for example Sandmo (1971). The main result (Sandmo) is that the risk-averse competitive firm will produce less than a risk-neutral competitive firm or one which knew with certainty that the price was going to be equal to its expected value. Drèze (1985)has used risk as a means of introducing a more realistic model of the firm into general equilibrium theory. General equilibrium theory is beyond the scope of this essay; but we should note that he does 'inquire within' and that his approach has much in common with that of Aoki (1984).

This brings us to a fork on the profit-maximizing branch. The divorce between ownership and control is explicitly recognized and the theory of agency developed to deal with it. The divorce occurs whenever an owner (or principal) submits a risky operation in which he has an interest to an operator (or agent) whose conduct he cannot monitor costlessly. Thus the theory of agency, originally developed in the discussion of sharecropping (risk sharing) and other forms of tenancy (see Stiglitz 1974) has the widest application, evidently to insurance, and, of particular interest in the present context, to the interior operations of firms, not only the relationship between owners and controllers but even between managers and teams (of employees) (see particularly Ross 1973; Jensen and Meckling 1976; Holmstrom 1982; Grossman and Hart 1983). It is commonly cheaper to give the operator (whether tenant, car-driver or executive) an incentive to good behaviour than to try to monitor him or her. This, of course, leads to less than optimal risk sharing (collision deductible in automobile insurance). Another incentive to good behaviour in the face of costly monitoring is suggested by Eaton and White (1983): this is to give an employee a bonus, a wage above his or her opportunity cost, so that, in the case that misconduct is detected, dismissal is a genuine penalty (see also Shapiro and Stiglitz 1984). Thus both carrots and sticks have been considered. When behaviour is unobservable, incentive compatibility may require some surprising forms of contract. Thus Holmstrom has shown that the only way to avoid the free-rider problem in a team in which effort is not observable is a contract which threatens to break the budget: deliver the target, or no member gets anything (someone else takes the full value of whatever is delivered). This raises two immediate problems. First, it may pay the 'someone else' to bribe a member of the team to shirk ('just a little'). Second, if achievement of the target depends on effort and some random variable (s), how would risk-averse members of the team dare to enter into such a contract?

Above I distinguished between two approaches to the theory of the firm, that of the maximizers and of those who wished to 'inquire within'. In agency theory we see the two converging. We are 'within', but not for its own sake; the agenda is still the allocation and distribution of scarce resources. We are forced within to deal, inter alia, with problems raised by Adam Smith over two centuries ago, in conjunction with our own better understanding of risk.

Let us now consider other optimizing models. These depend not merely upon the divorce between ownership and control but on the idea that there is 'slack' within which the controllers may play their own game without being noticed

and called to account. This in turn depends on the existence of market imperfections. The usual story has been that large firms are typically in a position to make monopoly rents, and that these rents can be forgone, used up, or ploughed back at the discretion of the controllers. It is acknowledged that rents usually turn out to be quasi-rents, but suggested that the large firms (conglomerates) can, by heavy R&D expenditure, enjoy a perpetual stream of quasi-rents: while one source is being competed away, another is being developed (perhaps patented). Thus there is always some room for discretionary expenditure by the controllers. This room may in turn be limited by the perspicacity of the capital market, but it is suggested (Marris 1964) that the power of the capital market to discipline controllers is limited by the costs of information and the fact that the supply of capital to potential takeover raiders is not infinitely elastic. Suppose, however, that capital markets were perfect. So long as the divorce between ownership and control remained, so would the problem of arranging incentive-compatible contracts for managers, whoever owned the equity.

How much scope for discretionary behaviour there actually is, then, is an empirical question to which we do not have a final answer. There is, however, no shortage of models of how managers will behave if they have the room – room to maximize their own utility functions, that is. We have Baumol (1959): maximize growth subject to a minimum profit constraint. Marris (1964) and J. Williamson (1966) offer more sophisticated versions. O.E. Williamson (1964) introduces the idea of 'expense preference'. The controllers can dissipate the rents by padding costs in ways which increase their utility. These ideas (and there are others) have obvious application to regulated industries, at least in the case in which the regulatory standard is a profit ceiling. Marris and J. Williamson both take into account the financial structure of the firm. There is now a large literature on this subject which I shall not discuss here.

(The first formal application of utility maximization to the theory of the firm was probably Scitovsky's 1943. I have not listed him above because I take him to be writing of a Smithian entrepreneur taking time off to play golf rather than following the 'divorce branch'.)

The set of 'other' models may be seen to subdivide again, between behaviourism, and something more purposeful associated with the work of Herbert Simon ('don't maximize, Simonize!'). To be sure, the firms in Cyert and March wanted to make a profit: they just do not seem to have been very good at it. Along the 'Simon branch' we have purposeful, self-interested behaviour. We may call it rational too, as long as it is understood that optimization is thought to be too difficult, and it is accordingly rational not to try. It does not follow that optimization does not occur: firms may adopt a convergent process, as in Day and Tinney (1968). In a 'sufficiently stable' environment, convergence might, of course, be quite common. But convergence must be proved rather than optimization assumed. It is thought rational for the firm to adopt routines or standard operating procedures that work at least 'well enough'. The meaning of 'innovation' is now extended. The introduction of a new routine that successfully handles a complicated decision that has to be taken with limited information is as much an innovation as a new product or an improvement in the technology. (From this point of view, a new legal or financial instrument that reduces transactions costs is an innovation too.)

It would not, I think, be a good use of space to catalogue all Simon's own innovations and suggestions. (For more recent discussion of bounded rationality, and related matters, see March 1978.) Instead, I shall consider only a recent contribution on this branch, the work of Nelson and Winter (1982) already referred to. These writers are much concerned with economic growth, perhaps less in static allocational problems. They inherit from Schumpeter, and Marshall, as well as Simon, and they name Cyert and March among their intellectual ancestors, as well as Alchian (1950).

Nelson and Winter argue that firms do not know the well-defined technological choice sets of standard theory. They only know how to do what they do do, and how to make at least local searches to do other things. Thus there is no sharp distinction between the choice set and the choice, and maximization is not an appropriate concept or

mode of analysis. Neither is equilibrium for either firm or industry. The configuration of an industry at any time is seen as the outcome of an evolutionary process, whence the appropriate tool is a Markov process (as in Newman and Wolfe 1961). The 'genes' required for biological analogy are the firms' routines: the standard procedures (in production, marketing, finance, and so on) that it knows how to operate. Its environment is stochastic, and the firm continually has to search for new routines (mutations). Chance enters twice. The search for a new routine may be deliberate, but its success is subject to chance. Once discovered, its application is subject to chance. Thus we have purposeful, self-interested behaviour, but success is a matter of luck. Routines are inherited, but new routines, once discovered, may also be copied by others, which allows the evolutionary process to be much faster than the biological process. There is another important point here. Nelson and Winter show that it may be more profitable to wait and to copy an innovation made by others than to incur the expenses necessary to develop it oneself. This seems to be contrary to the Schumpeterian intuition. There is also a shift in focus from the 'firm'. For Nelson and Winter the evolution of the industry is the subject of study, and the routines are the genes in the evolutionary process. The 'firm', although it is assumed to adopt purposeful, self-interested conduct (to seek profit), is not itself a matter of particular interest: it is something of a transient which happens, at any moment of time, to have inherited some routines, and may or may not succeed in developing some new, successful, ones. As in the earlier biological analogies, success will be rewarded and failure punished, but this is not advanced as an argument for 'as if' optimizing behaviour; it is part of the evolutionary process. Indeed, Nelson and Winter offer the first formal proof that, in this process, it is the profitable firms that survive. For other problems (R&D and technological change; Schumpeterian competition), they have to rely on simulation techniques which, however well handled, always leave one a little uncertain about what has been established, or, at least, at what level of generality.

It is now time to return to the question posed at the beginning of this article: what is the scope and purpose of the theory of the firm? Indeed, is there *a* theory of the firm at all? Perhaps not. There is a file of optimizing models. We may include in this file the theory of agency and much recent work on information and incentives. (There are also inquiries into such organizational matters as integration and the divisional structure of large corporations, which I do not discuss here.) In the 'other' branch, profit-seeking but not optimizing, there is the recent work by Nelson and Winter, in which the focus is on the development of the industry (population), and the firm is little more than an agent (unit organism) for the transmission of genes. And there is recent work, very exciting work, exploiting the ideas of capital commitment and credible threats, much of it in the spatial literature, on the strategic behaviour of firms in small group situations. Much of this work has been associated with developments in game theory. I shall not describe it here on the possibly dubious grounds that it is better filed as 'Industrial organization' or 'theory of market structure'. Demarcation lines are not, of course, well established; it could be argued that, whenever we invoke the ubiquitous Cournot-Nash equilibrium concept, we are taking a game-theoretic approach, and some might wish to interpret theory of the firm more widely than I have done. Be that as it may, there is clearly no such thing as *a* theory of the firm. But there is a great deal in the file, subdivide it as we will, and since the Second World War we have seen great advances, on many different fronts, albeit differently motivated and with different methodological orientations.

See Also

- ▶ Advertising
- ▶ Competition and Selection
- ▶ Corporations
- ▶ Entrepreneurship
- ▶ Ideal Output
- ▶ Marginal and Average Cost Pricing
- ▶ Market Structure
- ▶ Monopoly
- ▶ Oligopoly
- ▶ Predatory Pricing
- ▶ Price Discrimination (Theory)

Bibliography

Alchian, A.A. 1950. Uncertainty, evolution and economic theory. *Journal of Political Economy* 58: 211–212.

Alchian, A.A., and H. Demsetz. 1972. Production, information costs and economic organization. *American Economic Review* 62: 777–795.

Andrews, P.W.S. 1949. *Manufacturing business*. London: Macmillan.

Aoki, M. 1984. *The co-operative game theory of the firm*. Oxford: Clarendon Press.

Arrow, K.J. 1971. *Essays in the theory of risk-bearing*. Chicago: Markham.

Baumol, W.J. 1959. *Business behavior, value and growth*. New York: Macmillan.

Berle, A.A., and G.C. Means. 1933. *The modern corporation and private property*. New York: Macmillan.

Brechling, F.P.R. 1975. *Investment and employment decisions*. Manchester: Manchester University Press.

Chamberlin, E.H. 1933. *The theory of monopolistic competition*. Cambridge, MA/London: Harvard University Press/Oxford University Press.

Coase, R.H. 1937. The nature of the firm. *Economica* NS 4: 386–405.

Cournot, A.A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: L. Hachette. 1897 *Researches into the mathematical principles of the theory of wealth* (trans: Bacon, N.T.). London/New York: Macmillan.

Cyert, R.M., and J.G. March. 1963. *A behavioral theory of the firm*. Englewood Cliffs: Prentice Hall.

Darwin, C. 1859. *On the origin of species*. London: Murray.

Day, R.H., and E.H. Tinney. 1968. How to cooperate in business without really trying: a learning model of decentralized decision making. *Journal of Political Economy* 76: 583–600.

Dorfman, R., P.A. Samuelson, and R.M. Solow. 1958. *Linear programming and economic analysis*. New York: McGraw-Hill.

Drèze, J.H. 1985. Uncertainty and the firm in general equilibrium theory. *Economic Journal* 95 (Suppl): 1–20.

Eaton, B.C., and W.D. White. 1983. The economy of high wages: an agency problem. *Economica* NS 50: 175–182.

Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*, ed. M. Friedman. Chicago: University of Chicago Press.

Grossman, S.J., and O.D. Hart. 1983. An analysis of the principal-agent problem. *Econometrica* 51: 7–46.

Grossman, S.J., and J. Stiglitz. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70: 393–408.

Hall, R.L., and C.J. Hitch. 1939. Price theory and business behaviour. *Oxford Economic Papers* 2: 12–45.

Holmstrom, B. 1982. Moral hazard in teams. *Bell Journal of Economics* 13: 324–340.

Jensen, M.C., and W.H. Meckling. 1976. Theory of the firm: Managerial behaviour, agency costs and ownership structure. *Journal of Financial Economics* 3: 305–360.

Knight, F. 1921. *Risk, uncertainty and profit*. Boston: Houghton Mifflin.

Koopmans, T.C., ed. 1951. *Activity analysis of production and allocation*, Cowles commission monograph. Vol. 13. New York: Wiley.

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

Machlup, F. 1946. Marginal analysis and empirical research. *American Economic Review* 36: 519–554.

Machlup, F. 1967. Theories of the firm; marginalist, behavioral, managerial. *American Economic Review* 57: 1–33.

March, J.G. 1978. Bounded rationality, ambiguity, and the engineering of choice. *Bell Journal of Economics* 9: 587–610.

Marris, R. *1964*. *The economic theory of 'Managerial' Capitalism*. London: Macmillan.

Marshall, A. 1890. *Principles of economics*. London: Macmillan.

Mill, J.S. 1848. *Principles of political economy, with some of their applications to social philosophy*. London: J.W. Parker.

Mortensen, D.T. 1970. A theory of wage and employment dynamics. In *Microeconomic foundations of employment and inflation theory*, ed. E.S. Phelps. New York: W.W. Norton.

Nelson, R.R., and S.G. Winter. 1982. *An evolutionary theory of economic change*. Cambridge, MA/London: Harvard University Press.

von Neumann, J., and O. Morgenstem. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.

Newman, P. 1960. The erosion of Marshall's theory of value. *Quarterly Journal of Economics* 74: 587–601.

Newman, P., and J.N. Wolfe. 1961. A model for the long-run theory of value. *Review of Economic Studies* 29: 51–61.

Penrose, E.T. 1952. Biological analogies in the theory of the firm. *American Economic Review* 42: 804–819.

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

Ross, S. 1973. The economic theory of agency: The principal's problem. *American Economic Review* 63: 134–139.

Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.

Sandmo, A. 1971. On the theory of the competitive firm under price uncertainty. *American Economic Review* 61: 65–73.

Schumpeter, J.A. 1934. *The theory of economic development*. Cambridge, MA: Harvard University Press.

Scitovsky, T. 1943. A note on profit maximization and its implications. *Review of Economic Studies* 11: 57–60.

Shapiro, C., and J.E. Stiglitz. 1984. Equilibrium unemployment as a worker discipline device. *American Economic Review* 74: 433–444.

Simon, H.A. 1952. On the application of servomechanism theory in the study of production control. *Econometrica* 20: 247–268.
Simon, H.A. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics* 69: 99–118.
Simon, H.A. 1959. Theories of decision making in economics. *American Economic Review* 49: 253–283.
Simon, H.A. 1979. Rational decision making in business organizations. *American Economic Review* 69: 493–513.
Smith, A. 1776. *An enquiry into the nature and causes of the wealth of nations*. London: W. Strahan and T. Cadell.
Stigler, G.J. 1939. Production and distribution in the short run. *Journal of Political Economy* 47: 305–327.
Stiglitz, J.E. 1974. Incentives and risk sharing in sharecropping. *Review of Economic Studies* 41: 219–255.
Treadway, A.B.. 1969. On rational entrepreneurial behaviour and the demand for investment. *Review of Economic Studies* 36: 227–239.
Viner, J. 1931. Cost curves and supply curves. *Zeitschrift für Nationalökonomie* 3: 23–46.
Walras, L. 1874. *Eléments d' économie politique pure*. Lausanne/Paris/Basle: L. Corbaz/Guillaumin/H. Georg.
Whitin, T.M. 1953. *The theory of inventory management*. Princeton: Princeton University Press.
Williamson, O.E. 1964. *The economics of discretionary behavior: Managerial objectives in a theory of the firm*. Englewood Cliffs: Prentice-Hall.
Williamson, J. 1966. Profit, growth and sales maximization. *Economica* NS 33: 1–16.
Williamson, O.E. 1970. *Corporate control and business behavior*. Englewood Cliffs: Prentice-Hall.
Williamson, O.E. 1980. The organization of work: A comparative institutional assessment. *Journal of Economic Behavior and Organization* 1: 5–38.
Winter, S.G. 1975. Optimization and evolution in the theory of the firm. In *Adaptive economic models*, ed. R.H. Day and T. Groves. London/New York: Academic.

Firm-Level Employment Dynamics

Jeff Campbell

Abstract

Firm-level employment dynamics deals with the evolution of firms' employment decisions when they face costs of creating and destroying jobs. It lies at the intersection of labour economics, industrial organization and macroeconomics. Recent contributions use rational expectations models of labour demand to match salient statistics from establishment-level employment records.

Keywords

Adjustment costs; Firm-level employment dynamics; Job creation and destruction; Structured and unstructured jobs

JEL Classifications

D4; D10

Firm-level employment dynamics is the branch of economics that deals with the evolution of firms' employment decisions. The static analysis of labour demand equates a firm's marginal product of labour with the wage. Observation suggests that this abstracts from important considerations of employers when expanding or contracting their firms. Recruiting new employees requires effort, and preparing them for the jobs at hand might require training. Employees with substantial tenure might have legal rights that make their dismissal costly. All these realistic constraints make a firm's current employment complementary with its level at any future date. Hence, a firm's employment decisions when properly considered are *dynamic*. Firm-level employment dynamics is the area of economics that seeks to understand this decision using both theory and measurement. It lies at the intersection of industrial organization, labour economics and macroeconomics. It shares with labour economics a central concern with the employment relationship. Because firm entry and exit plays a substantial role in the evolution of total employment, it shares with industrial organization an interest in entrepreneurship. Firm-level costs of adjusting employment provide one potential source of persistence in economy-wide employment, so the area has contributed to the macroeconomics of business cycles.

Early theoretical treatments of the firm's dynamic employment choices came from the Ph.D. theses of Oi and Rosen. Oi (1962) coined the adjective 'quasi-fixed' to describe a factor of production that could be changed from its

previous value at a price. He considered the labour demand of a firm facing a constant wage and interest rate that must incur recruiting and training costs when expanding employment. Denote these with W, r, and τ. Then the firm's first-order condition for labour is $P \times f'(N) = W + r \times \tau$, where the production function holding other inputs constant is $f(\cdot)$, P is the output price, and N is the firm's employment choice. Oi noted two fundamental implications of this equation. First, the marginal product of labour generally exceeds the wage, so wage-setting institutions must support such a gap if the firm is to recover its investment in job creation. Second, unexpected permanent reductions in P leave N unchanged so long as the marginal product of labour remains above the wage. Rosen (1968) extended this by noting that adjusting an employee's hours worked generally costs less than adding or dismissing workers. Hence, fluctuations in hours worked should be more important for workers with high training and recruiting costs. He examined the employment decisions of regulated railways and found that they conform to this pattern.

Oi and Rosen intuitively saw many of the fundamental theoretical implications of imposing labour adjustment costs, but the first fully dynamic treatment of the firm's labour demand curve came from macroeconomics. Sargent (1978) considered a firm with a quadratic production function maximizing profit subject to quadratic costs of adjusting employment. Given a stochastic wage, W_t, he showed that the firm's optimal labour demand curve takes the form

$$N_t = (1-\rho)N_{t-1} - \theta \mathrm{E}_t\left[\sum_{j=0}^{\infty} \lambda^j W_{t+j}\right].$$

Here, N_t is the firm's employment at date t, E is the expectations operator, and ρ and λ are positive parameters that depend on the interest rate, the cost of adjustment, and the production function's concavity. Intuitively, N_{t-1} influences the profit-maximizing choice of N_t because it changes the cost of achieving any given level of employment. The complementarity between current and future employment makes N_t a function of the wage and its expected value at all future dates. This rule for N_t aggregates easily: if all firms follow it, then total employment does so as well. Sargent estimated this using US data on private employment and real wages. His model considered both straight-time and overtime employment. The quarterly data did not contain substantial evidence against the model, but the response of employment to real wage fluctuations in the model was much less than that measured with a vector autoregression.

With quadratic costs of adjustment, firms smooth their employment adjustments across time. This conflicts with the casual observation that firms' employment adjustment is *lumpy*, that is, they alternate between periods with very little or no employment adjustment and others with high rates of hiring or firing. Further credence to that view came from observations of firm employment collected by national statistical agencies. Using plant-level employment observations from the Dutch economy in 1988 and 1990, Hamermesh et al. (1996) showed that 28.3 per cent of firms kept employment constant over that two-year period. All these firms changed the *identities* of their employees. The average hiring rate for these firms was 11.3 per cent, so they apparently face costs of changing the jobs in the firm that are independent of the costs of changing the workers filling them.

With their book-length study of plant-level employment dynamics in the US manufacturing sector Davis et al. (1996) reinforced the conclusion that firm-level employment adjustment is lumpy. Their data came from the Longitudinal Research Database, an unbalanced panel of quarterly firm-level employment observations created from the surveys underlying the Annual Survey of Manufacturers and the Census of Manufacturing. These are confidential US Census records, but they may be used for approved projects that benefit the US Census at one of several regional census research data centres. Denote the employment of firm i in quarter t with N_{it}, and let $N_t = \sum_{i=1}^{M_t} N_{it}$ be the employment of the M_t firms with positive employment in quarter t. With these data, Davis, Haltiwanger and Schuh defined the job creation and destruction rates as

$$POS_t \equiv 2 \times \sum_{it} I\{N_{it} > N_{it-1}\} \frac{N_{it} - N_{it-1}}{N_t + N_{t-1}}$$

$$NEG_t \equiv 2 \times \sum_{it} I\{N_{it} < N_{it-1}\} \frac{N_{it-1} - N_{it}}{N_t + N_{t-1}}.$$

So defined, the difference between these two rates equals the rate of employment growth. These authors refer to their sum as employment reallocation.

The examination of these statistics from 1972: IV to 1988: IV yielded the following conclusions. (1) The rates of job creation and destruction are both very large relative to total employment changes. The average annual rates of job creation and destruction equalled 9.1 and 10.3 per cent. (2) The job creation and destruction rates of the population of young and middle-aged plants (less than ten years old) are much higher than those of their older counterparts. (3) Plants' employment changes are persistent. Some 70 per cent of newly created jobs last at least one year, and 80 per cent of newly destroyed jobs fail to reappear within a year. (4) Employment adjustment is lumpy. Two-thirds of job creation and destruction occurs at plants that adjust their employment by 25 per cent or more. (5) Employment drops in a recession because job destruction increases. Job creation is relatively acyclical.

Together, these facts have become the empirical touchstone for firm-level employment dynamics.

These facts motivated the creation of new models of firms' employment choices that incorporated adjustment costs and lent themselves to aggregation. A pair of papers by Campbell and Fisher (2000, 2004) develops one such model and applies it to explain the job creation and destruction facts of Davis, Haltiwanger, and Schuh. They begin with the labour demand problem of a single plant that produces a homogenous good for sale in a competitive market. The plant uses one factor of production, labour, that comes in fixed shift lengths. The per-period cost of employment measured in units of the output price is W_t, and let n_t denotet employment at this plant. The plant's output in period t is $z_t n_t^{\alpha}$, where z_t is the plant's idiosyncratic productivity term and $0 < \alpha < 1$. The wage follows a Markov chain over $\{W_l \leq W_h\}$ with transition probability p, and the idiosyncratic productivity shock follows a random walk with bounded innovation ε_t. The production function's strict concavity could arise from limits to a manager's effective span of control. When the plant changes its employment, it incurs adjustment costs that are proportional to the number of jobs created or destroyed. If employment at the plant expands, the cost per job created in units of lost output is τc, and the analogous cost per job destroyed is τd.

With these primitives, the profit maximization problem for a plant manager discounting future profits with β can be represented as a dynamic programming problem with initial states n_{t-1}, zt, and W_t. Its associated Bellman equation is $v(n_{t-1}, z_t, W_t) = \max_{n_t} z_t n_t^{\alpha} - W_t n_t - \tau(n_t, n_{t-1})(n_t - n_{t-1}) + \beta \mathrm{E}_t[v(n_t, z_{t+1}, W_{t+1})]$.

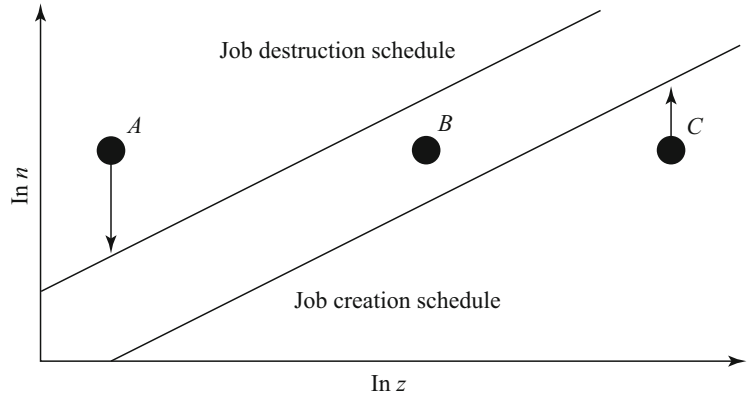
Here, $\tau(y, x) \equiv \tau_c \times I\{y > x\}$ -$\tau_d \times \{y < x\}$ is the per-job adjustment cost incurred. Campbell and Fisher (2000) show that the plant's optimal employment policy has a very simple structure. There exist job creation and destruction schedules, $\underline{n}(z, W) = \underline{y}(W) z^{1/(1-\alpha)}$ and $\overline{n}(z, W) = \overline{y}(W) z^{1/(1-\alpha)}$, such that

$$n_{t+1} = \begin{cases} \underline{n}(z, W) & \text{if } n_{t-1} \leq \underline{n}(z, W) \\ n_{t-1} & \text{if } \underline{n}(z, W) < n_{t-1} < \overline{n}(z, W) \\ \overline{n}(z, W) & \text{if } \overline{n}(z, W) \leq n_{t-1} \end{cases}.$$

Figure 1 illustrates these policies. On its horizontal axis is ln z_t, while its vertical axis gives ln n_{t-1} and ln n_t. The three plants labelled A, B, and C all start with identical values of n_{t-1} but different values of z_t. The job creation and destruction schedules are both linear with slopes equal to $1 = (1 - \alpha)$. Plant A lies above the job destruction schedule, so it reduces employment. Plant C lies below the job creation schedule, so it creates jobs. Plant B lies between the two schedules.

Here, the costs of job creation and destruction both exceed their associated benefits, so the plant's optimal employment is unchanged. Thus, this model automatically replicates one of Hammermesh, Hassink, and van Ours's findings: the plant's optimal employment frequently does not change.

F

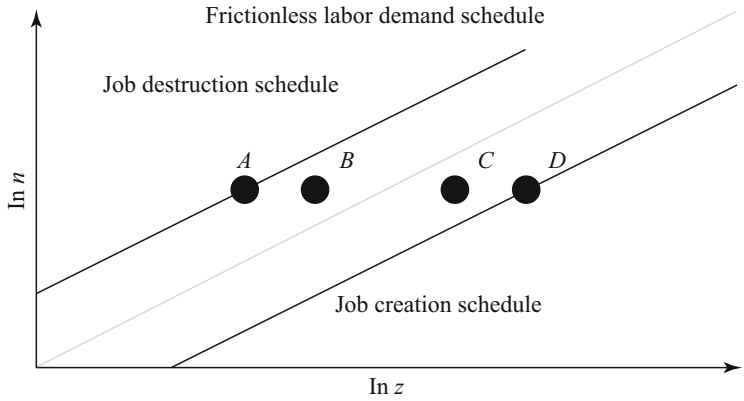


Firm-Level Employment Dynamics, Fig. 1 Optimal employment policy in Campbell and Fisher (2000)

Davis, Haltiwanger, and Schuh's finding that job destruction accounts for most cyclical employment variation attracted a great deal of attention in macroeconomics. Campbell and Fisher (2000) show that this simple model can replicate that fact if employment fluctuations arise from variation in W_t. To appreciate how this can be, note that the total cost of creating a job is $W_t + \tau_c$, which has an elasticity with respect to W_t that is less than one. The total cost of destroying a job is $W_t - \tau_d$, so its elasticity with respect to W_t exceeds 1. This asymmetry in the costs of job creation and destruction translates into asymmetric responses of the job creation and destruction schedules to changes in W_t. When the model is calibrated to match the characteristics of a typical US manufacturing industry, this microeconomic asymmetry produces the observed aggregate dynamics in a large population of such plants: the variance of job destruction exceeds the variance of job creation.

Campbell and Fisher (2004) extend this model to address Davis, Haltiwanger, and Schuh's finding that the magnitude of job creation and destruction declines with a plant's age and a related fact: *aggregate* fluctuations in young and middle-aged plants' employment exceed those of employment at older plants. To do so, they incorporate a life cycle into the above model. Plants exit exogenously and are instantly replaced by new entrants. All entrants begin life in a 'volatile' state with high probability of exit and high idiosyncratic productivity variance. In each period, a plant has a constant probability of transiting to a 'stable' state with lower exit probability and idiosyncratic productivity variance. Alone, this change would (mechanically) replicate the finding that young plants display greater job creation and destruction rates than their older counterparts. To generate young plants' greater business-cycle sensitivity, Campbell and Fisher add 'unstructured' jobs. Creating and destroying these jobs is costless, but for a worker to fill such a job is less productive than filling a structured job, which is costly to create and destroy.

Intuition suggests that a plant's use of unstructured jobs depends on its position in the life cycle. Young firms face high uncertainty about their future productivity and survival, so they find unstructured jobs more attractive than their older more predictable counterparts. This is indeed the case. In the calibrated version of the model that Campbell and Fisher use, firms in the 'mature' life-cycle stage never use unstructured jobs. Their employment dynamics qualitatively mimic those in the simpler model. In contrast, young plants' greater uncertainty induces them to create fewer structured jobs. This increases the marginal product of labour and thereby makes creating unstructured jobs more attractive. Figure 2 illustrates such young firms' employment choices. As in Fig. 1, the four plants labelled *A*, *B*, *C*, and *D* all have the same previous employment in *structured* jobs. Job creation and destruction schedules govern these plants' choices of structured jobs. These are the figure's solid lines. The dashed line gives



Firm-Level Employment Dynamics, Fig. 2 Employment choices when structured and unstructured jobs are used

the optimal employment in unstructured jobs if structured jobs were not available. Plants A and B do not use unstructured jobs, because they lie above this frictionless labour demand schedule. Plants C and D lie below it, and so they employ workers in both structured and unstructured jobs. Plants B and C both lie between the job creation and destruction schedules, so small changes in productivity induce neither of them to change their employment in structured jobs. However, only plant B would keep total employment constant. Plant C would change its employment in unstructured jobs following a small change in z_t. In this sense, the greater uncertainty young plants face leads them to choose more flexible production structures. Campbell and Fisher show in their calibrated version of this model that this greater microeconomic flexibility leads to larger aggregate responses to aggregate productivity shocks. Thus, the microeconomic differences between plants at different stages of the life cycle lead directly to the different aggregate differences in their employment dynamics.

One aspect of firm-level employment dynamics not captured by Campbell and Fisher's models is the prevalence of very large employment adjustments. To generate this, Bentolila and Bertola (1990) add fixed costs of employment adjustment. This non-convexity complicates the model's analysis, but under certain conditions a plant's optimal employment policy follows a two-sided version of an (S, s) policy familiar from inventory models. Denote the gap between a plant's actual employment and its optimal value without adjustment costs using gt. Then the firm lowers the gap to the target u by destroying jobs whenever it would otherwise exceed the trigger U, and it raises it to the target l by creating jobs whenever it would otherwise fall below the trigger L. Campbell and Fisher's model can be written in this form, where $u = U$ and $l = L$. Fixed costs of employment adjustment cause the targets to differ from their associated triggers and induce the firm to make only large employment adjustments.

Research on firm-level employment dynamics currently examines areas far removed from the initial focus on US manufacturing. Foote (1998) and Campbell and Lapham (2004) examine the dynamics of employment in service and retail industries. Foote finds that job creation dominates aggregate employment fluctuations in these industries. Consistent with this, Campbell and Lapham find that retail industries expand employment following a demand shock by increasing net entry. The importance of entrepreneurship for retail industries' employment is intuitive, and it suggests that the empirical and theoretical lessons learned from studying manufacturing industries will not apply easily to this important sector.

See Also

- ▶ Adjustment Costs
- ▶ Aggregation (production)

▶ Business Cycle measurement
▶ Firm Boundaries (Empirical Studies)
▶ Rosen, Sherwin (1938–2001)

Bibliography

Bentolila, G., and S. Bertola. 1990. Firing costs and labour demand: How bad is eurosclerosis? *Review of Economic Studies* 57: 381–402.
Campbell, J.R., and J.D.M. Fisher. 2000. Aggregate employment fluctuations with microeconomic asymmetries. *American Economic Review* 90: 1323–1345.
Campbell, J.R., and J.D.M. Fisher. 2004. Idiosyncratic risk and aggregate employment dynamics. *Review of Economic Dynamics* 7: 331–353.
Campbell, J.R., and B. Lapham. 2004. Real exchange rate fluctuations and the dynamics of retail trade industries on the U.S. – Canada border. *American Economic Review* 94: 1194–1206.
Davis, S.J., J.C. Haltiwanger, and S. Schuh. 1996. *Job creation and destruction*. Cambridge, MA: MIT Press.
Foote, C.L. 1998. Trend employment growth and the bunching of job creation and destruction. *Quarterly Journal of Economics* 113: 809–834.
Hamermesh, D.S., W.H.J. Hassink, and J.C. van Ours. 1996. Job turnover and labor turnover: a taxonomy of employment dynamics. *Annales d'Economie et de Statistique* 41(42): 21–40.
Oi, W. 1962. Labor as a quasi-fixed factor. *Journal of Political Economy* 70: 538–555.
Rosen, S. 1968. Short-run employment variation on class-I railroads in the U.S., 1947–1963. *Econometrica* 36: 511–529.
Sargent, T.J. 1978. Estimation of dynamic labor demand schedules under rational expectations. *Journal of Political Economy* 86: 1009–1044.

Fiscal and Monetary Policies in Developing Countries

David Fielding

Abstract

Low levels of economic development constrain fiscal and monetary policy in several ways. Few developing countries are able to raise much direct tax revenue, and so must rely on other sources of funding, including seigniorage. Institutional constraints often lead to a high risk of hyperinflation and currency crises. Credible, effective institutions can be created with appropriate outside help, but there are few examples of this in practice.

Keywords

Budget deficits; Central bank independence; Currency board; Currency crises; Developing countries; Development economics; Direct taxation; Exchange rate peg; Fiscal policy in developing countries; Foreign aid; Heterodox macroeconomics; Inflation targeting; Inflation; Laffer curve; Monetarism; Monetary policy in developing countries; Monetary unions; Phillips curve; Public expenditure; Seigniorage; Tariffs; Taylor rules; Time consistency

JEL Classifications

O2

Policymakers in developing (low-income, semi-industrialized) countries face particular challenges when setting taxes, interest rates and quantitative monetary instruments. All that follows should be preceded by a caveat: developing countries encompass at least as much economic diversity as the OECD. There is no such thing as a representative developing country; much harm can be (and has been) done by the incautious application of stylized models from development macroeconomics to individual countries. Nevertheless, we can identify those characteristics of developing countries that are likely to impose severe constraints on macroeconomic policymaking.

Fiscal and Monetary Characteristics of Developing Countries

The descriptive statistics in Table 1 provide some insight into the ways in which the fiscal and monetary characteristics of many developing countries differ from those of the developed world. The first row of the table contains information about fiscal structure and financial development in the United

Fiscal and Monetary Policies in Developing Countries, Table 1 Selected descriptive statistics for 2000

	Direct taxes (% of total taxes)	Import tax (% of total taxes)	Total taxes (% of GDP)	M1/M2 (%)	M2/ GDP (%)
United States	61	01	20	24	60
Countries with per capita GNI < $10K[a]					
Africa	27	33	19	55	33
Americas	21	10	13	23	29
Asia	16	11	09	26	30
Europe	11	02	16	30	17
Countries with per capita GNI < $5K[a]	21	21	15	47	34
Countries with per capita GNI of $5–10K[a]	24	11	20	37	37

[a]The *per capita* gross national income (GNI) figures are PPP-adjusted. The averages are constructed from those countries for which complete data are available in World Bank (2003): Algeria, Bolivia, Bulgaria, Congo Republic, Costa Rica, Cote d'Ivoire, Croatia, Dominican Republic, El Salvador, Estonia, Georgia, India, Iran, Jamaica, Jordan, Kazakhstan, Latvia, Lithuania, Madagascar, Mauritius, Mexico, Moldova, Mongolia, Nepal, Pakistan, Paraguay, Peru, Philippines, Poland, Romania, Russia, Sri Lanka, St. Vincent, Swaziland, Tajikistan, Thailand, Tunisia, Turkey, Uganda, Ukraine, Uruguay, Venezuela and Vietnam. Russia and Turkey are included in the figures for Europe.

States. Subsequent rows show equivalent average figures for those low-income countries for which data are available.

The table indicates some of the structural differences between the developing country average and the United States:

- Direct taxation in developing countries makes up a much smaller fraction of total tax revenue, and import duties make up a larger fraction.
- In Asia and the Americas, total tax revenue makes up a substantially smaller fraction of GDP.
- In Africa and Europe, M1 makes up a much larger fraction of M2.
- M2 makes up a much smaller fraction of GDP.

All these features are more pronounced for countries with a per capita gross national income below $5000 than for those countries in the $5000–$10,000 range.

The low levels of direct taxation in developing countries reflect that fact that a large fraction of private sector income is non-monetized: for example, many peasant households grow subsistence crops for their own consumption. Even when income is monetized, the administrative costs of direct taxation are often relatively high because of, for example, low levels of literacy and limited information technology. Governments are therefore forced to rely to a much greater degree on seigniorage revenue and on import duties. (High tariffs are often motivated by the need for fiscal revenue rather than by import substitution.) Inflation in developing countries is usually far higher than in the OECD. Between 1990 and 2000, the average annual inflation rate for the median developing country in Table 1 was 46 per cent; only five countries had single-digit inflation, and 14 had average inflation rates over 1000 per cent per annum.

The low levels of broad money demand in developing countries reflect low savings rates and limited access to financial services. Commercial banks are often absent from rural areas, where low per capita income, low population density and poor transport and communication infrastructure entail high costs in financial service provision to individual customers. In many developing countries a large fraction of the total money stock is in the form of cash, and few households have access to interest-bearing assets. One consequence is that the interest elasticities of saving and money demand are often very low; another is that there is limited scope for absorption of public debt by the domestic private sector.

Monetary Policy: The Neoclassical Perspective

Central to the monetarist approach to development macroeconomics is the argument that in developing countries high inflation is always and everywhere a fiscal phenomenon. Agénor and Montiel (1999) provide an extensive survey of this approach. In the standard formulation of the argument, which embodies many of the constraints highlighted above, there is a Cagan money demand function:

$$M/P = \exp\left(-\alpha \cdot \pi^{\beta}\right) \tag{1}$$

where M is the nominal money stock, P is the price index and $\pi = \dot{P}/P.M$ is to be interpreted as narrow money. There are no interest-bearing assets and no interest elasticity of money demand, so the opportunity cost of holding money depends just on the inflation rate. There is also a fixed real budget deficit, D, financed entirely be seigniorage:

$$D = \dot{M}/P = [M/P] \cdot \mu \tag{2}$$

where $\mu = \dot{M}/M$. This reflects the government's limited access to tax revenue and domestic credit. Combining Eqs. (1) and (2) we have:

$$D = \exp\left(-\alpha \cdot \pi^{\beta}\right) \cdot \mu \tag{3}$$

Equation (1) entails that an equilibrium with a constant π requires $\pi = \mu$, so that M/P is constant. For low enough values of D there will be two such equilibria, solutions to Eq. (3) with $\pi = \mu$. But for high values of D there is no equilibrium: successively higher levels of inflation lead to lower levels of real money demand, requiring higher rates of monetary expansion to finance the budget deficit, and so yet more inflation.

The first goal of macroeconomic policy is therefore to reduce the budget deficit to a level compatible with a stable inflation rate. In the absence of alternative sources of revenue this entails a reduction in public expenditure, which may have a negative impact on social and economic development. This provides a rationale for foreign aid to subsidize public expenditure in the medium term, while the country develops the institutions that will facilitate a wider fiscal base and a financial sector that will support some public debt. This approach still views the main macroeconomic function of a central bank in a developing country as generating seigniorage revenue. Policy reform is intended to reduce seigniorage, not to zero, but to a range compatible with a stable inflation rate. Indeed, a part of the neoclassical development macroeconomics literature analyses the inflation tax using concepts explicitly drawn from public finance, for example the Laffer Curve. The use of monetary policy for business cycle stabilization is at most a secondary objective.

Time Consistency in Monetary Policy

The critique of Kydland and Prescott (1977) can readily be applied to a seigniorage model. The simple model above provides an extreme case. Consider a policymaker for whom D is a variable to be maximized, subject to the equilibrium condition that $\pi = \mu$. From Eq. (3), the optimal rate of monetary expansion is $[\alpha \cdot \beta]^{-1/\beta}$; rates higher than this will reduce revenue. But if we modify Eq. (1) so that current money demand is based on a predetermined *expectation* of inflation, then for a given expectation and a given level of money demand the optimum inflation rate is infinite. The rational expectation of inflation is therefore infinite, in which case money demand and revenue are zero. The policymaker's problem is how to pre- commit credibly to a rate of expansion equal to $[\alpha \cdot \beta]^{-1/\beta}$. Failure to solve this problem is one suggested reason for the failure of disinflation programmes in developing countries.

The standard solution to the time inconsistency problem in industrialized countries is to delegate control of monetary policy to a central bank governor with a contract to target a given inflation rate. The constraint facing many developing countries is the absence of a political tradition or political institutions that will give people confidence in any laws enacted to create central bank

independence. Evidence on the link between central bank independence and inflation in developing countries is very weak. There is no significant correlation between historical inflation rates and historical indices of independence, which are based on the assumption that laws in developing countries have the same force as those in industrialized countries (Cuckierman et al. 1992). One interpretation of these results is that legislation for central bank independence would of little use in many developing countries, either because independence de jure does not entail independence de facto or because underdeveloped political institutions are unable to deliver enough clarity or stability in the decision-making process to allay people's doubts.

A possible alternative to central bank independence legislation is commitment to a credible nominal exchange rate peg. For a given real exchange rate a fixed peg against, for example, the euro or US dollar delivers an inflation rate equal to that of the eurozone or the USA. However, a fixed peg will be credible in the long run only if it is accompanied by an appropriate rate of domestic monetary expansion. Excessive domestic monetary expansion will lead to persistent balance of payments deficits and a loss of official foreign exchange reserves; eventually this will cause a collapse in the demand for domestic currency.

'First generation' currency crisis models show that with excessive monetary expansion this collapse can happen long before official reserves are finally depleted (Flood and Garber 1984). It would be irrational to hold on to domestic currency until reserves were finally depleted: at that point there would be a discrete fall in the value of domestic currency as the exchange rate shifted to a market value unsupported by central bank intervention, and those left holding domestic currency would make a loss. Instead, people will offload domestic currency as soon as monetary expansion has driven the implicit market value without intervention below the pegged rate.

'Second generation' models (Obstfeld 1996) go a step further, explaining currency crises in cases where there is moderate monetary growth, no greater than the rate of growth of the supply of foreign currency. Private sector views on the probability of an imminent abandonment of a peg will depend on an assessment of the likely costs and benefits of the peg for the government. (One example of such a scenario is when seigniorage revenue is higher under a more flexible exchange rate regime, but such flexibility deters foreign investment.) But these views will themselves influence the current level of demand for foreign and domestic currency, and so the opportunity cost of maintaining the peg. Models of such an environment typically imply the existence of multiple equilibria. There may be an equilibrium with a low perceived probability of collapse and a low opportunity cost of maintaining the peg, but this equilibrium is unlikely to be globally stable. The feedback between the perceived probability of collapse and the true opportunity cost of the peg means that some rumour questioning the government's commitment to the peg, however small and baseless, could eventually undermine this commitment, regardless of its fiscal and monetary discipline.

Table 2 illustrates some cases in which monetary and fiscal discipline has not been sufficient for the maintenance of an exchange rate peg. In the three cases shown, the size of the budget deficit in the years prior to collapse was not an excessive fraction of GDP (Bolivia, Honduras), or else seigniorage revenue did not account for a large fraction of the deficit (Zambia).

One interpretation of such examples is that they emphasize the need for monetary institutions free from all domestic political pressures and whose commitment to monetary discipline is without doubt. Lost seigniorage revenue is not the only cost of an exchange rate peg. An appreciation of the euro or US dollar due to idiosyncratic shocks in the eurozone or the USA is likely to create a recession in any country pegging to one of these currencies. These recessions can make the peg very unpopular. If we ignore the question of whether such unpopularity is justified, one suggested route to the creation of politically independent monetary institutions is the establishment of currency boards. In a currency board system the central bank is legally required to back issue of domestic currency one-for-one with reserves in a

Fiscal and Monetary Policies in Developing Countries, Table 2 Budget deficits and seigniorage revenue in the run-up to the abandonment of an exchange rate peg

	Bolivia (T = 1982)		Honduras (T = 1990)		Zambia (T = 1981)	
	Deficit/ GDP	Seigniorage/ deficit	Deficit/ GDP	Seigniorage/ deficit	Deficit/ GDP	Seigniorage/ deficit
T–3	0.074	0.151	0.036	0.415	0.144	0.040
T–2	0.079	0.401	0.030	0.336	0.091	0.058
T–1	0.204	0.240	0.033	0.525	0.185	0.051

Note: The peg in each country was abandoned in year T
Sources: IMF (1983, 1999)

given foreign currency. In eastern Europe currency boards have met with some success, at least in terms of maintaining a fixed peg. Recent examples are Bosnia–Herzegovina and Bulgaria pegging to the euro, and Latvia and Lithuania pegging to the US dollar. By contrast, the currency board system in Argentina met with spectacular failure, showing that currency board systems can be abandoned with almost as much ease as a conventional fixed peg.

A second suggested route to independent monetary institutions is the formation of monetary unions. A transnational central bank may well be free from many of the political pressures facing the central bank of a single country. In order to exert any political pressure on their central bank, the governments (and populations) of a monetary union would need to coordinate their actions. At any one time, conflicting economic interests are likely to undermine coordination attempts. It is always possible to secede from a monetary union, but in the absence of existing national monetary institutions this is potentially very costly. Given the ill will that secession is likely to generate among the remaining members of the union, it is also likely to be an irreversible decision, unlike the abandonment of a currency board. This irreversibility is likely to deter governments from abandoning their commitment to the monetary union. Currently, there are three major monetary unions among developing countries. These are the East Caribbean Currency Union (ECCU: Anguilla, Antigua, Dominica, Grenada, Montserrat, St Kitts, St Lucia, St Vincent), the West African Economic and Monetary Union (UEMOA: Benin, Burkina Faso, Côte d'Ivoire, Guinea–Bissau, Mali, Niger, Senegal, Togo) and the Economic and Monetary Community of Central Africa (CEMAC: Cameroon, Central African Republic, Chad, Congo Republic, Equatorial Guinea, Gabon). The ECCU has for decades maintained a fixed peg to the US dollar with a currency board arrangement. The two African monetary unions have maintained a fixed peg against the French franc (and now the euro) since the member states' independence in the 1960s, with just one devaluation in 1994. All three monetary unions have maintained low and stable rates of inflation.

However, it is unlikely that these three monetary unions could easily be replicated elsewhere. The ECCU is a group of small island economies where tourism makes up a large fraction of GDP and the US dollar circulates freely anyway; the monetary institutions just formalize pre-existing dollarization. The African monetary unions maintain a peg in cooperation with the French government. The French treasury exchanges euros for the two African currencies at a fixed rate, so the peg does not constrain the two central banks' use of domestic monetary instruments in the short run. (There are rules to prevent excessive monetary expansion in the long run.) Moreover, the French provide overdraft facilities to the two central banks to help cushion balance of payments shocks. So when the euro appreciates because of macroeconomic shocks specific to Europe, the African countries are not obliged to live through a recession. The feasibility of the peg has relied on an unusually strong (and arguably neo-colonial) economic commitment from the country issuing the anchor currency. Otherwise, it is likely that countries without credible domestic monetary institutions can buy a low inflation rate only at

the cost of a fixed peg that periodically generates damaging recessions. Even if 'second generation' currency crises can somehow be averted, weak domestic monetary institutions and incomplete information about policymakers' preferences mean that any relaxation of the exchange rate regime in times of recession will undermine the credibility of the commitment to low inflation.

Monetary Policy: Alternative Perspectives

There is a body of literature that encompasses alternatives to the monetarist approach discussed above. This literature is often labelled 'heterodox' in the context of policy formation and 'structuralist' in the context of theoretical models, of which Cardoso (1981) is a good example. At its core is the idea that inflationary spirals can be generated by the wage- and price-setting institutions within an imperfectly competitive economy, with the supply of money responding passively to increases in prices.

Suppose for example that industrial prices (p) in a closed economy are set by monopolistic firms as a mark-up on nominal industrial wages (w):

$$p = [1 + \theta].w \tag{4}$$

Workers would like to maintain a fixed real wage, so in equilibrium the ratio of nominal wages to consumer prices will be fixed. If consumption is made up of industrial goods and non-industrial goods in fixed proportions (φ, $1 - \varphi$), then the constant real wage condition can be written as:

$$w = \eta \cdot [\varphi \cdot p + (1 - \varphi) \cdot q] \tag{5}$$

where η is the target real wage and q is the price of non-industrial goods. (The closed economy and fixed consumption share assumptions are not essential to this class of model.) Together Eqs. (4) and (5) pin down relative prices:

$$p/q = \frac{1 - \varphi}{[\eta \cdot (1 + \theta)]^{-1} - \varphi} \tag{6}$$

There is a positive relationship between relative prices and the target real wage. Now if the supply of non-industrial goods depends just on relative prices, Eq. (6) will pin down non-industrial production and hence also, with full employment and a given level of resources, industrial production. In general, these production shares will not be equal to the consumption shares φ and $1 - \varphi$, in which case the model is overdetermined and has no equilibrium. An inflationary spiral will exist if non-industrial prices adjust to clear goods markets at a level of p/q less than $[1 - \varphi]/\{[\eta \cdot (1 + \theta)]^{-1} - \varphi\}$, which entails a real wage less than η. In such a world workers will raise wage demands, so $\dot{w}/w > 0$, but from Eq. (4) $\dot{q}/q = \dot{w}/w$ and with non-industrial prices adjusting to maintain the initial level of p/q we also have $\dot{q}/q = \dot{p}/p$ so there is no change in the real wage; nominal wages and prices will rise indefinitely.

Various policy prescriptions follow from such a model. A government-imposed nominal industrial wage freeze will halt the inflationary cycle at no cost to industrial workers, since their real wages are constant for all levels of inflation. Alternatively, subsidizing consumption of the non-industrial good will raise the real wage and reduce inflationary pressure. This is the macroeconomic basis for arguments in favour of food subsidies. One criticism of subsidies is that they increase the size of the budget deficit, so in models that integrate monetarist and structuralist elements the impact of subsidies on inflation is ambiguously signed.

Evidence on the effectiveness of heterodox anti-inflation measures, compared with orthodox fiscal and monetary contraction, is very limited. Many of the high-profile programmes designed to tackle hyperinflation in the 1980s (for example, Argentina and Israel in 1985, Brazil in 1986 and Mexico in 1987) combined fiscal and monetary reforms with heterodox wage and price controls of one kind or another. It is very unclear which elements of these programmes were crucial in determining their success or failure. There is some limited evidence from reduced- form macro-econometric models on the direction of causality between wage growth, price growth and money growth (for example, Montiel 1989).

F

This suggests that different macroeconomic processes are at work in different countries. On the basis of current evidence, policy prescriptions for any one country should be accompanied by a large caveat.

Taylor Rules in Developing Countries

The discussions above relate to the problems developing countries face in achieving a stable fiscal policy environment with a moderate rate of monetary growth. This has been the main focus of the theoretical and empirical literature to date. However, there is also a growing literature that extends the mainstream concerns of the monetary policy literature in OECD countries – in particular, issues surrounding the optimal policy response to exogenous macroeconomic shocks – to developing countries.

Certainly, developing countries are at least as vulnerable to external shocks as OECD countries. Many developing countries are small in size, trading a relatively large fraction of their GDP and exporting a narrow range of primary commodities for which world prices are highly volatile. In these countries, yearly changes in the terms of trade can increase or reduce domestic income by several percentage points. The average value of such changes over 1990–2000 in the developing countries in Table 1 is greater than three per cent of GDP; in the USA it is less than 0.2 per cent. Values in excess of ten per cent have been recorded for some countries in some years. These figures are large relative to the magnitude of supply shocks estimated for most OECD countries. So there is a strong case for advocating an active short- run monetary policy in those developing countries with a stable underlying monetary and fiscal regime.

The current norm in OECD countries is an institutionally independent central bank which regularly adjusts a monetary policy instrument – the quantity of short-term lending to commercial banks, or more frequently the corresponding interest rate – in order to meet an implicit or explicit medium-term inflation target. This target is usually accompanied by an injunction to avoid 'unnecessary' volatility in real macroeconomic indicators such as GDP growth or the unemployment rate. The relative weight to be given to the two goals is seldom explicit, but the academic literature – including the research divisions of many central banks, though never their policy statements – interprets the trade-off between output and price stability in terms of the framework introduced by Taylor (1993). That is, the optimal value of the instrument in any one period is derived from the maximization of an objective function including the deviations of inflation and output (or unemployment) from their target values, subject to a constraint embodied in a short-run supply curve (or Phillips curve). Shocks to the supply curve shift the constraint and so change the optimal value of the instrument; the magnitude of the change depends on the weights on the different targets in the objective function. Past central bank behaviour is often interpreted as such a Taylor rule plus inertia reflecting model uncertainty and a random component reflecting unquantifiable information about the economy.

In recent years, some non-OECD countries have introduced explicit inflation targeting with a degree of central bank independence and accountability. These are not countries typical of those in the study of Cuckierman et al. (1992): they have relatively stable political institutions and relatively democratic governments. High-profile examples are Brazil, Chile, the Czech Republic, Poland and South Africa. There are also central banks that lack an explicit inflation target but nevertheless publish policy reports that motivate the adjustment of monetary instruments by reference to short-term movements in inflation and output, accompanied by research division working papers on the application of Taylor rules to their economy. The central banks of the two African monetary unions discussed above, the BCEAO and the BEAC, are examples of this phenomenon. Econometric studies of the evolution of monetary instruments and macroeconomic variables in these counties suggest that in most cases their monetary institutions function in a broadly similar way to those of OECD countries, although the shocks to which they are responding (often normalized in econometric analysis!) are greater in magnitude.

The creation of such institutions is surely endogenous to a country's level of political and economic development. They are the consequence rather than the cause of a stable policy environment and relatively developed financial markets. The former ensures the credibility of monetary institutions; the latter ensures an identifiable monetary transmission mechanism in which interest rate changes can be expected to impact on the economy in a consistent way. Such examples represent one tail of the distribution of institutional quality, in which institutions reduce macroeconomic instability. There are still many countries in which institutions increase macroeconomic instability, as witnessed by the hyperinflation still endemic in many parts of the world. Nevertheless, they indicate that a high level of per capita GDP is not a necessary condition for monetary institutions equally as effective as those in the OECD.

See Also

- ▶ Development Economics
- ▶ Exchange Rate Volatility
- ▶ Hyperinflation
- ▶ Inflation Targeting
- ▶ International Financial Institutions (IFIs)
- ▶ Taylor Rules

Bibliography

Agénor, P.-R., and P. Montiel. *Development macroeconomics,* 2nd ed. Princeton: Princeton University Press.

Cardoso, E. 1981. Food supply and inflation. *Journal of Development Economics* 8: 269–284.

Cukierman, A., S. Webb, and B. Neyapti. 1992. Measuring the independence of central banks and its effect on policy outcomes. *World Bank Economic Review* 6: 353–398.

Flood, R., and P. Garber. 1984. Collapsing exchange rate regimes: Some linear examples. *Journal of International Economics* 17: 1–13.

IMF (International Monetary Fund). 1983. *International financial statistics yearbook*. Washington, DC: IMF.

IMF (International Monetary Fund). 1999. *International financial statistics yearbook*. Washington, DC: IMF.

Kydland, F., and E. Prescott. 1977. Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy* 85: 473–490.

Montiel, P. 1989. Empirical analysis of high inflation episodes in Argentina, Brazil and Israel. *IMF Staff Papers* 36: 527–549.

Obstfeld, M. 1996. Models of currency crises with self-fulfilling features. *European Economic Review* 40: 1037–1047.

Taylor, J. 1993. Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy* 39: 195–214.

World Bank. 2003. *World bank indicators 2003*. Washington, DC: World Bank.

F

Fiscal Federalism

David E. Wildasin

Abstract

Fiscal federalism is concerned with the division of policy responsibilities among different levels of government and with the fiscal interactions among these governments. Public service provision by lower-level governments can be efficiency-enhancing, although competition for mobile resources can also interfere with efficient resource allocation both in the public and private sectors. Intergovernmental transfers affect the overall equity and efficiency properties of public policies. Global economic integration and political and economic reforms in developing and transition economies – which have institutional contexts very different from those of the mature federations – present important challenges for a 'second generation' of federalism research.

Keywords

Decentralization; Factor mobility; Fiscal competition; Fiscal federalism; Horizontal equity; Intergovernmental grants; Local public goods; Optimal currency areas; Planning; Policy coordination; Stabilization policy; Tax competition; Tax distortions; Tiebout hypothesis

JEL Classifications

H3

Fiscal federalism is concerned with the division of policy responsibilities among different levels of government and with the fiscal interactions among these governments.

The Institutional Context of Fiscal Federalism

Fiscal federalism has long been a topic of keen interest in the United States and Canada. In both nations, subnational governments have traditionally played major roles in the provision of important public services, notably in the areas of education, health, social services, transportation, public safety, and economic development. In addition to non-tax revenues, subnational governments in both countries have had significant sources of tax revenues, with state/provincial governments relying heavily on retail sales taxes and taxes on personal and business income and with local governments depending on property taxes. Higher-level governments (national in relation to subnational, and state/provincial in relation to local) have supported the finances of lower-level governments with extensive programmes of intergovernmental fiscal transfers in order to promote the provision of particular public goods and services, to supplement (or perhaps displace) lower-level government taxes, and to advance broad social welfare objectives. Although they are subject to constitutional, statutory, and regulatory constraints, state/provincial and local governments exercise substantial fiscal autonomy with respect to expenditures, taxation and borrowing. National and subnational fiscal policies have been developed and implemented within the context of continuously evolving but fundamentally durable market, political, and legal institutions, underpinned by stable democratic constitutional structures.

There are long-established federations (and long traditions of scholarly research on federalism) in other parts of the world as well, but interest in fiscal federalism has become particularly intense in developing and transition economies since the early 1990s, no doubt in part because of broad reform initiatives that have reduced the role of the state in economic planning and control (Wildasin 1997a, ch. 2). In many of these countries, constitutional, economic, and political reforms have led to significant decentralization of tax, expenditure, and borrowing responsibilities, often accompanied by the development of new systems of intergovernmental fiscal transfers. In contrast to the mature North American federations, the newly (or increasingly) decentralized and liberalized economic and fiscal systems of many developing and transition economies are being implemented in the absence of the background political, legal, and market institutions found in more developed nations. The development and restructuring of federations around the world has presented many practical challenges and, for scholars, important questions regarding the design of federal systems, the implementation of fiscal reforms in such systems, and the interactions between basic social institutions and the public sector in federations.

Fiscal federalism is also a subject of increased interest and concern in the European Union. Fiscal decentralization has accompanied economic and political reforms in several European nations. In addition, the interactions of tax, expenditure, debt, and monetary policies among EU member states continuously raise questions concerning international policy coordination and the development of EU-wide supranational institutions. Controversy surrounds the issues of national sovereignty and the upward transfer of powers from national governments to EU executive, legislative, and judicial bodies. In important respects, however, the EU can be viewed as an emerging federation in which EU-level political and fiscal institutions are gradually developing within the context of an increasingly integrated and expanding system of developed and transition economies. From this perspective, the EU itself is a (so far relatively limited) higher-level government in relation to the national governments of its member states.

Fiscal federalism is thus a subject of great interest throughout the world. Wide international variation in the institutional context of federalism has stimulated what Oates (2005) calls a 'second generation' of fiscal federalism research, differentiated

from 'first-generation' research by its heightened attention to political, constitutional, financial and macroeconomic institutions. For example, issues of fiscal discipline, soft budget constraints, and subnational government borrowing, little discussed within the context of traditional federalism research, have received considerable attention in recent years (Inman 2003; Wildasin 1997b, 2004), especially with reference to newly decentralizing fiscal systems. Because the policy issues and institutional context of federalism varies widely throughout the world, a rapidly growing literature deals with fiscal federalism in an international context, often focusing on unique policy issues facing individual countries (see, for example, Bird and Vaillancourt 1998; Martinez-Vasquez and Alm 2003; and Rodden et al. 2003, which contain many studies of federalism problems in developing and transition economies).

As the foregoing remarks suggest, problems of fiscal federalism touch upon almost all aspects of fiscal policy, in almost all nations (especially the large nations and economic regions) of the world. The subject is correspondingly very broad. The following paragraphs highlight recurring themes that have occupied researchers for many years as well as selected issues that are likely to be the subject of active enquiry in coming years. The discussion begins with fundamental issues regarding the economic functions of different levels of government, noting their implications for the organization of the public sector. The potential efficiency gains from decentralized policymaking as well as the limitations of decentralization are discussed next, emphasizing the importance of resource mobility and fiscal competition as a crucial feature of the decision-making environment facing lower-level governments. Finally, directions for new research are briefly discussed.

The Organization of the Public Sector

What economic functions can, do, or should be performed by different levels of government? This fundamental question has been a focus of the federalism literature from its inception. There has been a broad normative consensus (Oates 1972) that, of Musgrave's (1959) 'three branches of the public household', the highest-level government (normally a national government, but possibly a supranational entity like the EU) should take responsibility for stabilization functions (that is, macroeconomic and monetary policies), that allocative functions (the provision of public goods and services and correction of market failures) should be undertaken by governments whose jurisdictional boundaries are co-terminous with the geographical scope of the regions affected by these policies, and that higher-level governments should be responsible for policies that target the distribution of income. Subnational economies are comparatively more open than national economies, which means that the impacts of stabilization policies are diluted through capital, labour, and financial flows when undertaken by lower-level governments; see, for example, Mundell's (1961) classic work on optimal currency areas. Similarly, the mobility of labour and capital constrains the ability of (small, open) subnational governments to alter the net distribution of income. For example, high taxes on the rich in one jurisdiction create incentives for the rich to locate elsewhere, while the provision of generous cash or in-kind benefits for the poor attracts beneficiaries (Stigler 1957). In addition to distorting the efficiency of resource allocation, the spatial reallocation of resources in response to local redistributive policies limits the set of feasible policies as well as their impact on net incomes. Lower-level governments may, however, serve effectively to provide public goods and services in the amounts that are most efficiently adapted to local benefits and costs, which normally vary among locations in accordance with differences in demographic composition, incomes, and technologies (Oates's 'decentralization theorem').

Allocative Efficiency at the Local Level

The decentralization theorem shows that non-uniform provision of public goods, varying in accordance with local benefits and costs, may be more efficient than uniform provision.

In principle, however, an omniscient and omnipotent central planner could implement optimal non-uniform policies, obviating the need for distinct administrative units of lower-level government. Such a planner could manage all public sector functions (in fact, all economic decisions) for the entire world. A key idea in the literature of fiscal federalism, however, is that lower-level units of government may be better informed about and more responsive to local demands. The *information* needed for efficient decision-making, and the *incentives* to use this information, may differ by level of government, just as markets provide incentives guiding decentralized market decisions for households and firms in ways not achievable, in practice, by central planning mechanisms.

This idea is developed explicitly, if informally, in Tiebout (1956). Tiebout draws the analogy between consumers shopping for commodities in the marketplace and households choosing residences from among a collection of localities. Writing soon after and in response to Samuelson's classic contributions to public goods theory, Tiebout asserts that households reveal their preferences for local public goods when they choose where to reside. Different localities provide different levels of public services, as illustrated by local school districts in the United States that offer different qualities of elementary and secondary education. Households with high valuations for education can outbid others for residences in localities with good schools, thus leading to a sorting of households by demand for public services. According to Tiebout, this matching of demand and supply leads to efficient provision of local public goods.

Tiebout's paper identifies local governments as distinct economic units that can perform important allocative functions in ways that central governments cannot. Tiebout is not specific, however, about exactly how local decision-makers determine public goods levels – whether by voting or through some other mechanism. Many subsequent contributions (see, for example, Wildasin 1986, for a survey and references), including both theoretical and empirical analyses, explore in detail the phenomenon of 'Tiebout sorting' and the implications of community stratification, by income, race, religion, age and other household attributes, for variation in local public expenditures. Median voter models (and variants thereof) commonly provide a theoretical starting point for empirical analyses of the demand for local public goods. Linkages between housing markets and local fiscal policies, as revealed by hedonic price relationships, suggest that local voters have incentives to support policies that preserve property values. In the extreme, these linkages may obviate altogether the need for households to participate in the collective decision-making process, by providing profit-maximizing property developers and other market participants with the information and incentives to make efficient policy choices, resulting in completely market-driven provision of public goods (Fischel 2001, discusses land use regulation, property development and their interactions with community formation and local policymaking).

In addition to the information and incentives that may result from the mobility of households and firms, emphasized by Tiebout, decentralized policymaking may also provide a framework for experimentation and learning about policy alternatives and their consequences as well as for learning about the performance of policymakers themselves (Besley and Case 1995).

Limits to Decentralization: Efficiency and Distributional Considerations

Tiebout's analysis and much subsequent research highlights the potential benefits, especially with respect to the efficiency of public good provision, from competition among lower-level governments for mobile households and firms. The potential disadvantages of fiscal decentralization have long been recognized, however. For instance, the economic service areas for local public goods may not closely match jurisdictional boundaries. Local health, educational, or transportation policies may benefit residents of neighbouring localities or society at large, spillover benefits that local decision-makers may ignore. These externalities can potentially be

internalized through voluntary policy coordination among neighbouring governments. Such coordination can be costly, however, resulting in inefficient decentralized public good provision. Within a federation, a higher-level government can use intergovernmental grants (generally conditional grants, especially matching grants that reduce the marginal cost of public good provision for recipient governments) in order to induce more efficient provision of externality-generating local public goods and services (Breton 1965). If the spillover benefits from a public good are sufficiently widespread, a higher-level government may assume complete responsibility for its provision. Such centralization of a governmental function involves a trade-off between the potential benefits from internalization of externalities and the potential informational disadvantages of centralized collective decision-making for a larger and more heterogeneous population (Alesina and Spolaore 2003).

A second possible drawback of decentralized policymaking arises if there are significant limitations on the fiscal instruments available to lower-level governments. In the competition among lower-level governments for households and businesses, taxes (or non-tax revenue instruments such as user fees or licenses) perform a 'price like' function by rationing access to public services. Taxes may also introduce inefficiencies of their own, however, not only through 'classical' tax distortions (distortion of in *situ* labour/leisure, consumption, savings, and investment decisions) but more especially through their effects on the locational choices of households and businesses. For example, subnational government income taxes may inefficiently drive profitable businesses and high-income households into low-tax jurisdictions, and retail sales taxes may encourage inefficient cross-boundary shopping. Fiscal competition for mobile factors of production or consumers may discourage taxation of these resources, changing the composition of the subnational revenue structures toward less-mobile tax bases if these are available and potentially constraining the overall level of government revenues. Underprovision of public goods may result, which, as in the case of spillover benefits, may potentially be remediated with well-designed fiscal transfers from higher-level governments (Wildasin 2006a; Wilson and Wildasin 2004; Wilson 1999). On the other hand, if Leviathan governments are likely to engage in excessive spending, fiscal competition may impose useful constraints on their revenue-raising powers (Brennan and Buchanan 1980).

A further difficulty for federalized systems arises from the fact that many public policies, by their nature, intermingle allocative and distributional impacts, so that a clean separation of allocative and redistributive functions between higher- and lower- level governments may be unattainable. Health, education, transport, economic development, and many social services involve allocative functions (service delivery for geographically limited areas) but also promote distributional goals. Particularly when competition among lower-level governments results in the formation of communities that are relatively homogeneous (with respect to income, race, age or other socioeconomic characteristics), the efficiency gains from decentralization may be realized in part precisely through increased disparities in public service provision. The demand for education, for example, is a normal good, so that stratification of localities by income produces disparities in educational quality between rich and poor localities, as efficiency requires. In the United States, concern about the fairness of inequality in education, partly as expressed in state government constitutions, has resulted in extensive litigation leading to judicial mandates for policy reforms, notably including extensive programmes of equalizing fiscal transfers from state to local governments (Inman and Rubinfeld 1979). More generally, the equalization of fiscal transfers from higher- to lower-level governments provides a mechanism through which to limit horizontal inequities in the fiscal treatment of households in rich and poor jurisdictions and the locational incentives to which they give rise (Boadway and Flatters 1983).

As noted earlier, factor mobility imposes constraints on the ability of governments to redistribute incomes. The integration of capital and labour markets can improve the efficiency of factor allocations and thus raise output and welfare, an important potential benefit that underpins policy initiatives,

such as economic integration within the EU, that seek to remove barriers to factor mobility. Factor mobility also affects factor prices, giving rise to potentially important first-order distributional impacts. Thus, economic integration affects not only the cost of 'decentralized' redistribution – which, in a global context with international factor mobility, includes redistribution by national as well as subnational governments. By affecting factor prices and the underlying distribution of income, it also may increase or decrease the benefits of redistributive policies. International capital mobility and the migration of younger workers (both skilled and unskilled) from developing and transition economies to aging developed nations thus create new policy trade-offs, particularly for the extensive redistributive systems of North America and Western Europe (Wildasin 2006b), the consequences of which will unfold in coming decades.

Directions for Future Research

As noted at the outset, the challenges of policy and institutional reform throughout the world have stimulated new interest in fiscal federalism. The incentives embedded in the institutional structures of the mature federations seem to have ensured that subnational governments maintain sufficient fiscal discipline to avoid major widespread or recurring fiscal crises, while preserving their ability to exercise significant policy autonomy with respect to the level and composition of their taxes, expenditures and debts (Buettner and Wildasin 2006; Inman 2003; Wildasin 2004). Such institutions cannot be taken for granted, however, and many informed observers see potential risks from fiscal decentralization in the evolving federations of the developing and transition economies, including risks from excessive (that is, inefficiently high) spending or borrowing by subnational governments. An appropriate mix of revenue and expenditure assignments, intergovernmental fiscal transfers, borrowing flexibility, and policy autonomy is needed in order to realize the potential efficiency gains from fiscal decentralization (McLure and Martinez-Vasquez n.d.; Weingast 2006). The interplay between the market environment (especially financial markets and institutions and capital and labour mobility), the assignment of fiscal and regulatory authorities by level of government, and the constraints that influence political decision-making is not well understood and promises to be the subject of extensive study in coming years.

The integration of national and international markets for labour and capital, of crucial importance for federalism, appears to be increasing over time, and affects the competitive pressures facing governments at all levels. The global configuration of age-imbalanced demographic structures (*y*oung poor populations in developing countries and old rich populations in developed countries) implies that international migration incentives are unlikely to diminish in the foreseeable future. The fiscal systems of developed nations, with their extensive systems of intra- and intergenerational transfers, will face growing challenges in coming decades as a result of population aging, even as competition for capital investment and mobile high-income households may increasing constrain their capacity to finance redistribution (Wildasin 2006c). Policy coordination, perhaps through newly developed governmental structures (for example, at the EU level), may provide opportunities for national governments to limit the degree of fiscal competition, helping them to finance the liabilities arising under existing redistributive systems. Alternatively, or in addition, national governments may explicitly or implicitly shift some expenditure responsibilities to lower-level governments as they manage growing fiscal imbalances arising from demographic change. In any case, growing fiscal imbalances are likely to form the backdrop for public finance in developed countries in coming decades, offering opportunities for fruitful analysis of the dynamics of factor mobility, factor market integration, dynamic fiscal adjustment, and institutional change within and among nations.

See Also

► Intergovernmental Grants
► Local Public Finance

- ▶ Public Finance
- ▶ Tax Competition
- ▶ Tiebout Hypothesis

Bibliography

Alesina, A., and E. Spolaore. 2003. *The size of nations*. Cambridge, MA: MIT Press.

Besley, T., and A. Case. 1995. Incumbent behavior: Vote seeking, tax setting and yardstick competition. *American Economic Review* 85: 25–45.

Bird, R.M., and F. Vaillancourt. 1998. *Fiscal decentralization in developing countries*. Cambridge: Cambridge University Press.

Boadway, R.W., and F. Flatters. 1983. Efficiency and equalization payments in a federal system of government: a synthesis and extension of recent results. *Canadian Journal of Economics* 15: 613–633.

Brennan, G., and J.M. Buchanan. 1980. *The power to tax*. Cambridge: Cambridge University Press.

Breton, A. 1965. A theory of government grants. *Canadian Journal of Economics and Political Science* 31: 175–187.

Buettner, T., and D.E. Wildasin. 2006. The dynamics of municipal fiscal adjustment. *Journal of Public Economics* 90: 1115–1132.

Fischel, W.A. *2001. The homevoter hypothesis: How home values influence local government taxation, school finance, and land-use policies*. *Cambridge, MA*: *Harvard University Press*.

Inman, R.P. 2003. Transfers and bailouts: Enforcing local fiscal discipline with lessons from US federalism. In Rodden, Eskeland and Litvack (2003).

Inman, R.P., and D.L. Rubinfeld. 1979. The judicial pursuit of local fiscal equity. *Harvard Law Review* 92: 1662–1750.

Martinez-Vasquez, J., and J. Alm. 2003. *Public finance in developing and transition countries: Essays in honor of Richard Bird*. Cheltenham: Edward Elgar.

McLure, C.E., and J. Martinez-Vasquez. n.d. *The assignment of revenues and expenditures in intergovernmental fiscal relations*. Washington, DC: World Bank.

Mundell, R.A. 1961. A theory of optimum currency areas. *American Economic Review* 51: 509–517.

Musgrave, R.A. 1959. *Theory of public finance*. New York: Macmillan.

Oates, W.E. 1972. *Fiscal federalism*. New York: Harcourt Brace Jovanovich.

Oates, W.E. 2005. Toward a second-generation theory of fiscal federalism. *International Tax and Public Finance* 12: 349–373.

Rodden, J., G.S. Eskeland, and J. Litvack. 2003. *Fiscal decentralization and the challenge of hard budget constraints*. Cambridge, MA: MIT Press.

Stigler, G.J. 1957. The tenable range of functions of local government. Joint Economic Committee. In *Federal expenditure policy for economic growth and stability*. Washington, DC: US Government Printing Office.

Tiebout, C.M. 1956. A pure theory of local expenditures. *Journal of Political Economy* 64: 416–424.

Wallack, J., and T.N. Srinivasan. 2006. *Federalism and economic reform: International perspectives*. Cambridge: Cambridge University Press.

Weingast, B. 2006. Second generation fiscal federalism: Implications for decentralized democratic governance and economic development. Unpublished paper presented at IFIR-CESifo Conference, New Directions in Fiscal Federalism. Lexington.

Wildasin, D.E. 1986. *Urban public finance*. New York: Harwood Academic.

Wildasin, D.E. 1997a. *Fiscal aspects of evolving federations*. Cambridge: Cambridge University Press.

Wildasin, D.E. 1997b. Externalities and bailouts: Hard and soft budget constraints in intergovernmental fiscal relations. Policy research working paper No. 1843. Washington, DC: World Bank.

Wildasin, D.E. 2004. The institutions of federalism: toward an analytical framework. *National Tax Journal* 62: 247–272.

Wildasin, D.E. 2006a. Fiscal competition. In *Oxford handbook of political economy*, ed. B. Weingast and D. Wittman. Oxford: Oxford University Press.

Wildasin, D.E. 2006b. Global competition for mobile resources: implications for equity, efficiency, and political economy. *CESifo Economic Studies* 52: 61–111.

Wildasin, D.E. 2006c. Public finance in an era of global demographic change: fertility busts, migration booms, and public policy. Unpublished paper presented at Council on Foreign Relations conference, Skilled migration today: Prospects, problems, and policies. New York.

Wilson, J.D. 1999. Theories of tax competition. *National Tax Journal* 52: 296–315.

Wilson, J.D., and D.E. Wildasin. 2004. Capital tax competition: Bane or boon? *Journal of Public Economics* 88: 1065–1091.

Fiscal Multipliers

Menzie Chinn

Abstract

The concept of fiscal multipliers is examined in the context of the major theoretical approaches. Differing methods of calculating multipliers are then recounted (structural equations, VAR, simulation). The sensitivity of estimates

to conditioning on the state of the economy (slack, financial system) and policy regimes (exchange rate system, monetary policy reaction function) is discussed.

Keywords

Crowding out; Debt; Deficits; Fiscal; Monetary reaction function; Portfolio balance; Ricardian equivalence; Spending; Taxes

JEL Classification

E62; E43; F33; F41

Introduction

The fiscal multiplier plays a central role in macroeconomic theory; at its simplest level, it is the change in output for a change in a fiscal policy instrument. For instance,

$$\frac{dY_t}{dZ_t}$$

where Y is output (or some other activity variable) and Z is a fiscal instrument, either government spending on goods and services, on government transfers, or taxes or tax rates. Since there are typically lags in the effects, one should distinguish between impact multipliers (above) and the cumulative multiplier:

$$\frac{\sum_{j=0}^{n} dY_{t+j}}{\sum_{j=0}^{n} dZ_{t+j}}$$

The interpretation of the fiscal multiplier is complicated by the fact that it is not a structural parameter. Rather, in most relevant contexts, the multiplier is a function of structural parameters and policy reaction parameters.

The issue of fiscal multipliers took on heightened importance in the wake of the 2008 global financial crisis, in which monetary policy and nondiscretionary fiscal policy proved insufficient to stem the sharp drop in income and employment. Substantial confusion regarding the nature and magnitude of fiscal multipliers arose; many of the disagreements remain.

This survey reviews the theoretical bases for the fiscal multiplier in differing frameworks. Then the differing methodologies for assessing the magnitude of differing multipliers are reviewed. Special cases and allowances for asymmetric effects are examined.

Theory

The Neoclassical Synthesis

The simplest way to understand multipliers is to consider an aggregate supply–aggregate demand model in the Neoclassical Synthesis – essentially a framework with short run Keynesian-type attributes and long run Classical properties. While this framework is not particularly rigorous, it turns out that many of the basic insights gleaned in other approaches can be understood in this framework.

For the moment, think of the aggregate demand as separable from the aggregate supply. Demand depends on fiscal policy and monetary policy, while the long run aggregate supply curve is determined by the level of technology, labour force, and capital stock. In the short run, a higher price level is associated with a higher economic activity.

Over time, the price level adjusts toward the expected price level and any deviation of output from full employment is eventually eroded. Hence, in the long run, the Classical model holds, so that any fiscal policy has zero effect. This framework is sometimes called the Neoclassical synthesis.

The more responsive the price level to the output gap, the smaller the change in income for any given government spending increase. In the extreme case, where there is no response of wages and prices to tightness in the labour and product markets, then the multiplier is relatively large. In this Keynesian model, the multiplier is a positive function of the marginal propensity to consume. From the national income accounting perspective,

a distinction has to be made between spending on goods and services, and transfer expenditures. The former will have a larger impact on output than the latter.

In the other extreme case, where wages and prices are infinitely responsive to the output gap, the short run aggregate supply and long run aggregate supply curve are the same. Then clearly the fiscal multiplier is zero. (Note that the supply side perspective can be interpreted in the framework of the Neoclassical Synthesis. The long run aggregate supply depends on the capital stock and labour force employed, as well as the level of technology. If marginal tax rate reductions increase employment and/or investment, then the multiplier for tax rate changes could be positive, even in the absence of demand effects.)

In addition, the multiplier also depends critically on the conduct of monetary policy. When policy controls the money supply, the multiplier depends on the income and interest sensitivities of money demand. In the more general case where there is a monetary policy reaction function, the multiplier will depend on the reaction function parameters. For instance, if the central bank is completely accommodative (i.e. keeps the interest rate constant), the multiplier is larger than if it is non-accommodative (as discussed further in the section on monetary regimes).

New Classical Approaches

The real business cycle (RBC) approaches can be thought of as stochastic versions of the Classical Models. One of the defining features of these types of models is the incorporation of microfoundations, in particular intertemporal considerations. With infinitely lived agents and no nominal rigidities, non-distortionary taxes have no impact on the present value of income. Hence, tax cuts have no impact on consumption, and thus on income. This tax cut result is often characterised as Ricardian equivalence (Barro 1974).

The implications of government spending are more difficult to analyse. In particular, if government spending is financed by higher non-distortionary taxes, then after tax income declines. As a consequence, labour effort increases, and output (measured as the sum of private and public consumption) rises. In the standard setup, where government consumption yields no utility, social welfare decreases even thought output rises.

When distortionary taxes are used to pay for government spending, then both output and social welfare will decline. Then the government spending multiplier would be negative.

While the stereotype of the RBC approach is consistent with small multipliers, small variations in the assumptions can deliver large multipliers. For instance, assuming that government capital and private capital and labour are complements can deliver large fiscal multipliers (Baxter and King 1993). Notice, however, that the multipliers in this case do not arise from the familiar demand-side effects, but rather from supply-side effects.

New Keynesian Models

New Keynesian models represent the result of combining microfounded models incorporating intertemporal optimisation with Keynesian-type nominal and real rigidities. Such models are associated with Gali and Woodford, for instance. The basis of these models are the real business cycle models, with money introduced using money in utility functions. The deviations from the RBCs usually come in the form of rigidities, both nominal and real. Nominal rigidities are often introduced by way of sticky prices; prices adjust at random points in time (often called Calvo pricing). Real rigidities often include adjustment costs (say, for investment) and deviations from full intertemporal optimisation: for instance, rule-of-thumb or hand-to-mouth consumers (e.g. Gali et al. 2007). In addition to allowing the models to fit the data better, the inclusion of these rigidities provides a role for fiscal as well as monetary policy.

Because the models are built around an essentially neoclassical framework, policies do not have large long run effects. However, in the short term, monetary and fiscal policies have an effect on output. The magnitude of the impact depends on the various parameters of the model, and – as in the Keynesian model – the nature of the monetary policy reaction function. An excellent overview of how these factors come into play in determining the multiplier is provided by Woodford (2011).

One key limitation highlighted by the financial crisis and the ensuing recession and recovery is the omission of financial frictions. In fact, the financial sector in the typical New Keynesian model is usually very simple (a single bond, for instance; in two-country models, uncovered interest parity might be relaxed by the inclusion of an *ad hoc* risk premium term).

Summing up, one can see that the different types of model will deliver fiscal multipliers of almost any magnitude. Moreover, even models of a particular class can deliver quite different multiplier values, depending on underlying parameter values and the assumptions regarding monetary policy reaction functions. As a consequence, one can only address the magnitude of multipliers by empirics.

Empirics

There are many ways of calculating multipliers, with the approaches often associated with certain theoretical frameworks. However, in general, there are three major approaches: (1) structural econometric, *à la* Cowles Commission; (2) vector autoregressions (VARs); and (3) simulation results from dynamic stochastic general equilibrium (DSGE) models. There are also other miscellaneous regression approaches.

Structural Econometric Approaches

The earliest approach to estimating multipliers involved estimating behavioural equations for the economy. Since the multiplier depends critically upon the marginal propensity to consume, estimates of the consumption function are central to the enterprise of calculating the multiplier. This enterprise is closely associated with the Cowles Commission approach to econometrics, which used (Keynesian) theory to achieve identification in multi-equation systems.

Large-scale macroeconometric models are the descendent of the early Keynesian Klein Goldberger model (Goldberger 1959), and – despite the disdain with which such models are held in academic circles – they still provide the basis for most estimates of multipliers. It appears that business sector economists still find such models useful for forecasting and policy analysis. They include the models run by Global Insight-IHS and Macroeconomic Advisers.

The equations in such models include, for instance, a consumption equation, an investment equation and price adjustment equations. Identification would require that there should be sufficient number of exogenous variables. Two assaults on this approach include the Lucas econometric policy evaluation critique, and the charge of incredible identifying assumptions (Sims 1980).

In the former case, the relevant question is whether the estimation procedure (which typically incorporates a complicated lag structure) actually identifies parameters that are invariant to policy changes (such as government spending changes). (Ericsson and Irons (1995) have argued that the Lucas critique is actually seldom relevant, given that large policy changes are rare.) In the latter, the concern is that identification is not possible, since there are very few truly exogenous variables. This concern motivates the enterprise of estimating vector autoregressions (described below).

While it is customary to disparage these types of model as eschewing intertemporal considerations, this characterisation is not always accurate. Some macroeconometric models incorporate model-consistent expectations – essentially an implementable version of rational expectations. Taylor (1993) is an early example of a relatively conventional macroeconometric model with forward-looking expectations. Other cases include the IMF's Multimod and the Fed's FRB/US model: see Laxton et al. (1998) and Brayton et al. (1997).

Vector Autoregressions (VARs)

Sims (1980) argued that the Cowles Commission approach to estimating large systems of equations required 'incredible' identifying assumptions. His alternative approach involved estimating a small system of equations, where each variable is modelled as a function of lags of all variables in the system. In Sims' original formulation, a recursive ordering is assumed.

Since there are no exogenous variables, the response is expressed in terms of the error term – or shock. That is, the response is expressed

in terms of the unpredictable component of government spending or tax revenues, and not in terms of a given change in either of those instruments.

There is no reason why the nature of shocks should follow a recursive ordering. Alternative approaches include long run restrictions, wherein one variable is not affected by a shock in another variable in the long run. This approach was pioneered in Blanchard and Quah (1989). Short run restrictions can also be incorporated, such that a shock to one variable has no immediate impact on another, as in Clarida and Gali (1994). Blanchard and Perotti (2002) used institutional features to add additional restrictions. Yet other types of restriction, including negative or positive responses, are also feasible (Mountford and Uhlig 2009). Ramey (2011b) focused on news in defence spending as a means of circumventing issues of identifying exogenous shocks. In all these cases, belief in the results depends upon how plausible one finds the identifying restrictions – including the restrictions on the number of relevant equations. These VARs typically employ relatively few equations, due to the large number of parameters that have to be estimated.

Another way of dealing with the issue of distinguishing between endogenous and exogenous fiscal measures is to use a narrative approach, as pioneered by Romer and Romer (1989) for monetary policy. Romer and Romer (2010) estimated the impact of tax changes on output using this approach.

Simulations Using Dynamic Stochastic General Equilibrium Models

In response to the criticism of the *ad hoc* nature of the large-scale macroeconometric models, most recent analyses of policy effects have been conducted using dynamic stochastic general equilibrium (DSGE) models which incorporate, to a greater or lesser degree, New Keynesian formulations.

The equations in these models are either calibrated (that is parameter values are selected) or estimated, or a combination thereof is used. The majority of these models incorporate Ricardian equivalence, contrary to the bulk of empirical evidence. Hence, almost by assumption, fiscal multipliers are typically small relative to those obtained in traditional macroeconometric models. In cases where Ricardian equivalence is dispensed with, multipliers are typically larger. (See for instance Kumhof et al. (2010). Note that instead of the future tax burden rising with spending, future spending might be restrained. Corsetti et al. (2010) and Corsetti et al. (forthcoming) trace out the dynamics in this case.)

Miscellaneous Approaches

Since multipliers are changes in output for a change in a fiscal instrument, estimation can proceed in a variety of ways. The simplest entails regression of output changes on instrument changes; the challenge is controlling for other effects. Since discretionary fiscal policy reacts, by definition, to other factors that might be unobservable to the econometrician, there are serious challenges to this approach.

For instance, Almunia et al. (2010) use panel regression analysis (in addition to VARs) for a set of countries; Nakamura and Steinsson (2011) for a set of states; and Acconcia et al. (2013) for Italian provinces. In contrast, Barro and Redlick (2009) use a long time series for the USA. (Reichling and Whalen (2012) survey 'local multipliers', which tend to focus on employment – rather than output – effects in subnational units. Other relevant studies (typically focusing on employment effects) include Chodorow-Reich et al. (forthcoming), Mendel (2012) and Moretti (2010).)

A Survey of Basic Results

Obviously the literature is too voluminous to review comprehensively. I focus first on the US-A. CBO (2012a, Table 2) has provided a range of estimates that the CBO considers plausible, based upon a variety of empirical and theoretical approaches (see Table 1).

For goods and services, the range is 0.5–2.5; in line with demand side models, the cumulative multiplier for government spending on transfers to individuals are typically lower, and range from 0.4 to 2.1. Tax cuts for individuals have a multiplier of between 0.3 and 1.5, if aimed at

Fiscal Multipliers, Table 1 Ranges for US cumulative output multipliers

Type of activity	Estimated output multipliers	
	Low estimate	High estimate
Purchase of goods and services by the Federal Government	0.5	2.5
Transfer payments to state and local governments for infrastructure	0.4	2.2
Transfer payments to state and local governments for other purposes	0.4	1.8
Transfer payments to individuals	0.4	2.1
One-time payments to retirees	0.2	1.0
Two-year tax cuts for lower- and middle-income people	0.3	1.5
One-year tax cut for higher-income people	0.1	0.6

Source: CBO (2012a, Table 2)

households with a relatively high marginal propensity to consume. (See the survey of approaches in the appendix to CBO (2012a).)

When assessing whether a government spending multiplier is large or small, the value of unity is often taken as a threshold. From the demand side perspective, when the spending multiplier is greater than one, then the private components of GDP rise along with government spending on goods and services; less than one, and some private components of demand are crowded out. (Since transfers affect output indirectly through consumption, multipliers for government transfers to individuals should be smaller than multipliers for spending on goods and services.)

Reichling and Whalen (2012) discuss the range of multiplier estimates associated with various approaches. Ramey (2011a) also surveys the literature, and concludes spending multipliers range from 0.8 to 1.5. Romer (2011) cites a higher range of estimates, conditioned on those relevant to post-2008 conditions.

The above estimates pertain to the US. Obviously, one can expand the sample to other countries and other times. Van Brusselen (2009) and Spilimbergo et al. (2009) survey a variety of developed country multiplier estimates.

Almunia et al. (2010) find, using a variety of econometric methodologies, that fiscal multipliers during the interwar years are in excess of unity, when looking across countries. Barro and Redlick (2009) incorporate WWII data in their analysis of US multipliers; critics have noted that rationing during the WWII period makes extrapolation of their results to peacetime conditions questionable.

Distinctions

Large, Closed vs. Small, Open Economies

Theory suggests that, at least from the demand side, fiscal multipliers should be smaller in open economies (where openness is measured in the context of trade of goods and services), holding all else constant. This is because the leakage from a small open economy due to imports or purchases of internationally tradable goods more generally rising with income mitigates the recirculation of spending in the economy. In a closed economy, the marginal propensity to import is arguably smaller. Ilzetzki et al. (forthcoming) estimate panel VARs and find that indeed small open economies have smaller multipliers.

In addition, for large economies, some portion of the leakage of spending that occurs through imports would return as increased demand for exports. That means that the large country multiplier would be larger than that for a small country, holding all other characteristics – such as trade openness – constant.

Fixed vs. Flexible Exchange Rate Regimes

Ilzetzki et al. find that countries under fixed exchange rates have larger multipliers than those under flexible exchange rates. This finding is in accord with the Mundell–Fleming model, which predicts that under fixed exchange rates, the monetary authority is forced to accommodate fiscal policy. With high capital mobility (which is likely in the set of countries examined), monetary policy has to be very accommodative, in order to maintain the exchange rate peg. Corsetti et al. (2012) obtain similar results regarding the magnitude of the multiplier, even after controlling for other factors (debt levels etc.) despite the fact that they

find the policy rate rises. They argue imperfect peg credibility accounts for this effect.

In a slightly different context, Nakamura and Steinsson (2011) confirm this result. Examining states in the USA, they find that the fiscal multiplier is 1.5 for government spending on goods and services. Since the USA is a monetary union, they interpret this multiplier as one pertaining to small economies on fixed exchange rates.

Monetary Regimes (Inflation Targeting, Zero Interest Rate Bound)

Perhaps the most important insight arising from the debates over fiscal policy during and after the great recession is that the multiplier depends critically on the conduct of monetary policy. This insight is obvious if one thinks about policy in a standard IS-LM framework, where the interest rate is constant either because of accommodative monetary policy (Davig and Leeper 2009), or because the economy is in a liquidity trap. Christiano et al. (2011) provide a rationale for this effect in the context of a liquidity trap in a DSGE.

Coenen et al. (2012) show that in DSGEs, the degree of monetary accommodation is critical. When central banks follow a Taylor Rule or inflation forecast-based rules, then multipliers are relatively small. However, when monetary policy is accommodative – that is interest rates are kept constant – then the cumulative multiplier is greater. This finding is consistent with the idea that fiscal policy in a liquidity trap is equivalent to a helicopter drop. As DeLong (2010) notes, when the price level is fixed, a helicopter drop changes nominal demand one-for-one, and therefore must have real effects. However, a helicopter drop is a combination of (i) an open market operation (OMO) purchasing bonds for cash, and (ii) a bond-financed tax cut. The monetary effects of an OMO plus the fiscal effects of a tax cut must therefore add up to the effects of a helicopter drop. In a liquidity trap, where one believes an OMO is powerless, fiscal expansion must therefore be powerful.

This insight is of particular importance because estimates of multipliers based upon historical data are likely to be less relevant in current circumstances, where interest rates have been kept near zero since 2008.

There is some evidence that the effects of fiscal policy in Europe have been unusually large in recent years (see Blanchard and Leigh forthcoming). One of the reasons is that the zero lower bound has prevented central banks from cutting interest rates to offset the negative short-term effects of fiscal consolidation.

Asymmetric Fiscal Effects

Many of the earlier studies assumed that the impact of fiscal policy was homogeneous across different states of the economy. Recent work has sought to relax this assumption. Given that the size of the multiplier is more relevant in certain circumstances than others, accounting for heterogeneous effects is critically important.

State-Dependent Multipliers

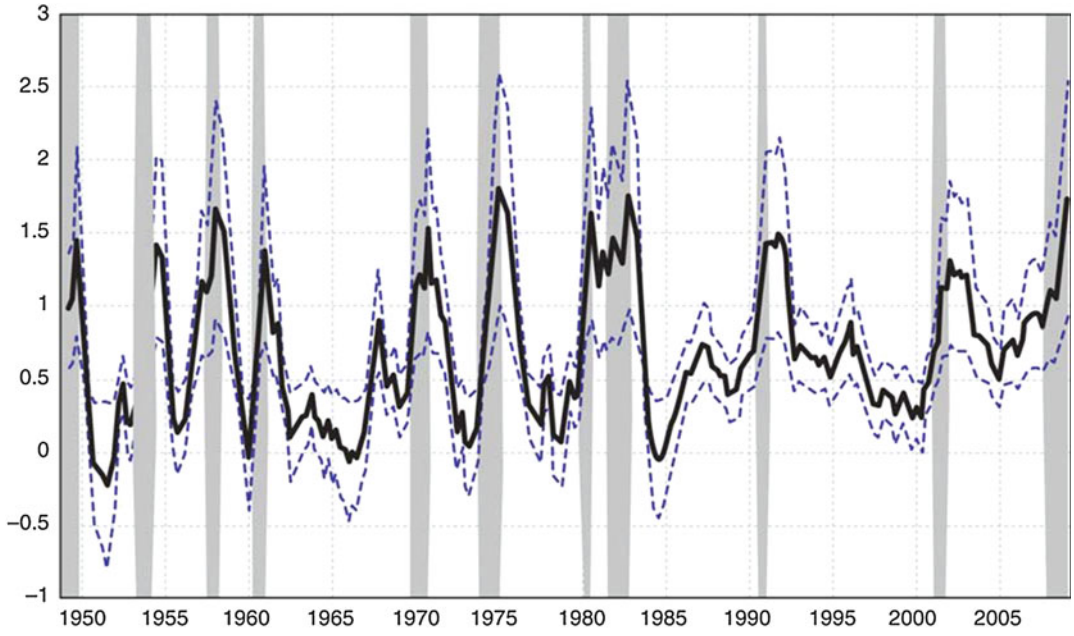
The demand side interpretation of the multiplier relies upon the possibility that additional factors of production will be drawn into use as demand rises. If factors of production are constrained, or are relatively more constrained, as economic slack disappears, then one might entertain asymmetry in the multiplier.

Auerbach and Gorodnichenko (2012a, b) and Fazzari et al. (2012) use VARs which allow the parameters to vary over expansions and contractions (Auerbach and Gorodnichenko use a smooth transition threshold where the threshold is selected a priori. Fazzari et al. estimate a discrete threshold.). Baum et al. (2012) condition on the output gap. The common finding in these instances is that multipliers are substantially larger during recessions.

To highlight the variation in the multiplier for the USA, I reproduce Figure 5 from Auerbach and Gorodnichenko (2012b), which plots their estimates of the multiplier over time (Fig. 1).

A different perspective on why long term multipliers are larger during periods of slack is delivered by Delong and Summers (2012). (Quantification of long-term impacts of depressed activity on potential GDP can be found in CBO (2012b).) They argue that long periods of depressed output can itself affect potential GDP, following the analysis of Blanchard and Summers

F



Fiscal Multipliers, Fig. 1 Historical multiplier for total government spending (Source: Auerbach and Gorodnichenko (2012b))

(1986). The prevalence of high rates of long-term unemployed is one obvious channel by which hysteretic effects can be imparted. When combined with an accommodative monetary policy or liquidity trap, the long term multiplier can be substantially larger than the impact multiplier.

Hence fiscal multipliers are largest exactly at times when expansionary fiscal policy is most needed. Estimates of multipliers based on averaging over periods of high and low activity are therefore useful, but not necessarily always relevant to the policy debate at hand.

Low Versus High Debt Levels

Ilzetzki et al. (forthcoming) determine that fiscal multipliers are essentially zero when debt is above (the sample) average. Corsetti et al. (2012) also find multipliers are smaller when public debt is high, controlling for other factors, although the measured differences are modest.

In high-debt situations, contractionary fiscal policy can in principle stimulate activity in the short run if it raises confidence in the government's solvency and reduces the need for disruptive adjustments later on (Blanchard 1990). A recent theoretical analysis of fiscal policy under conditions of high sovereign risk is by Corsetti et al. (forthcoming). A number of empirical studies find evidence of such expansionary effects (Giavazzi and Pagano 1990; Alesina and Perotti 1995; Alesina and Ardagna 2010; and others). Other papers suggest that such findings of expansionary effects are sensitive to how fiscal consolidation is defined (IMF 2010), and that the famous cases of expansionary contractions were typically driven by external demand rather than confidence effects (Perotti 2011).

Ordinary Versus Stressed Financial Systems

Historical estimates of the fiscal multiplier also condition on data when the financial system is operating normally, or is at least not highly impaired. However, the financial conditions during the crisis were arguably abnormal. To the extent that credit constraints were more binding (e.g. Eggertsson and Krugman 2012), households could be expected to behave in a more 'Keynesian' fashion, with less reference to 'permanent

income'. This would tend to result in a larger multiplier. See also Fernández-Villaverde (2010).

Corsetti et al. (2012), confirm empirically (using VARs) that during times of financial crisis, fiscal multipliers are larger. They conjecture that liquidity-constrained households are more pervasive during crises. They add the caveat that this finding holds true when public finances are strong.

Conclusion

The magnitude of the fiscal multiplier, in theory and in the data, depends on the characteristics of the economy. In some senses this observation is obvious. What is less recognised is that the state of the economy is as, or more, important than many other aspects that have been the focus of analysis. The most critical aspects include the degree of slack in the economy, the state of the financial system, and the conduct of monetary policy.

See Also

- ▶ Monetary and Fiscal Policy overviEw
- ▶ Neoclassical Synthesis
- ▶ New Keynesian Macroeconomics
- ▶ Vector Autoregressions

Acknowledgments I thank Giancarlo Corsetti, Brad DeLong, Jeffrey Frankel, Ethan Ilzetzki, Daniel Leigh, Felix Reichling, Carlos Vegh and the editor Garett Jones, for very helpful comments.

Bibliography

Acconcia, A., G. Corsetti, and S. Simonelli. 2013. Mafia and public spending: Evidence on the fiscal multiplier from a quasi-experiment. *Mimeo* (January).

Alesina, A., and S. Ardagna. 2010. Large changes in fiscal policy: Taxes versus spending. In *Tax policy and the economy*, vol. 24, ed. J.R. Brown. Cambridge, MA: National Bureau of Economic Research.

Alesina, A., and R. Perotti. 1995. Fiscal expansions and fiscal adjustments in OECD countries. *Economic Policy* 10(21): 205–248.

Almunia, M., A.S. Bénétrix, B. Eichengreen, K.H. O'Rourke, and G. Rua. 2010. From great depression to great credit crisis: Similarities, differences and lessons. *Economic Policy* 25(62): 219–265.

Auerbach, A.J., and Y. Gorodnichenko. 2012a. Fiscal multipliers in recession and expansion. In *Fiscal policy after the financial crisis*, ed. A. Alesina and F. Giavazzi. Chicago: University of Chicago Press.

Auerbach, A.J., and Y. Gorodnichenko. 2012b. Measuring the output responses to fiscal policy. *American Economic Journal: Economic Policy* 4: 1–27.

Barro, R.J. 1974. Are bonds net wealth? *Journal of Political Economy* 82(6): 1095–1117.

Barro, R.J., and C.J. Redlick. 2009. *Macroeconomic effects of government purchases and taxes*, NBER Working Paper, No. 15369 (September). Cambridge, Mass: National Bureau of Economic Research.

Baum, A., M. Poplawski-Ribeiro, and A. Weber. 2012. *Fiscal multipliers and the state of the economy*, IMF Working Paper No.12/286 (December). Washington, DC: International Monetary Fund.

Baxter, M., and R.G. King. 1993. Fiscal policy in general equilibrium. *American Economic Review* 83(3): 315–334.

Blanchard, O.J. 1990. Comment on Francesco Giavazzi and Marco Pagano, 'Can severe fiscal consolidations be expansionary? Tales of two small European countries. *NBER Macroeconomics Annual* 5: 111–116.

Blanchard, O., and D. Leigh. Forthcoming. Growth forecast errors and fiscal multipliers. *American Economic Review: Papers and Proceedings*.

Blanchard, O., and R. Perotti. 2002. An empirical characterization of the dynamic effects of changes in government spending and taxes on output. *Quarterly Journal of Economics* 117(4): 1329–1368.

Blanchard, O., and D. Quah. 1989. The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79(4): 655–673.

Blanchard, O., and L. Summers. 1986. Hysteresis and the European unemployment problem. *NBER Macroeconomics Annual* 1: 15.

Brayton, F., E. Mauskopf, D. Reifschneider, P. Tinsley, and J. Williams. 1997. The role of expectations in the FRB/US macroeconomic model. *Federal Reserve Bulletin (April)* 83: 227.

CBO. 2012a. *Estimated impact of the American recovery and reinvestment act on employment and economic output from October 2011 through December 2011*. Washington, DC: CBO.

CBO. 2012b. *What accounts for the slow growth of the economy after the recession?* Washington, DC: CBO.

Chodorow-Reich, G., L. Feiveson, Z. Liscow, and W.G. Woolston. Forthcoming. Does state fiscal relief during recessions increase employment? Evidence from the American Recovery and Reinvestment Act. *American Economic Journal: Economic Policy* 4: 118.

Christiano, L., M. Eichenbaum, and S. Rebelo. 2011. When is the government spending multiplier large? *Journal of Political Economy* 119(1): 78–121.

Clarida, R., and J. Gali. 1994. Sources of real exchange-rate fluctuations: How important are nominal shocks?

F

Carnegie-Rochester Conference Series on Public Policy 41: 1–56.
Coenen, G., et al. 2012. Effects of fiscal stimulus in structural models. *American Economic Journal: Macroeconomics* 4(1): 22–68.
Corsetti, G., K. Kuester, A. Meier, and G.J. Mueller. 2010. Debt consolidation and fiscal stabilization of deep recessions. *American Economic Review: Papers and Proceedings* 100(2): 41–45.
Corsetti, G., A. Meier, and G. Müller. 2012. What determines government spending multipliers? *Economic Policy* 27: 521–565.
Corsetti, G., K. Kuester, A. Meier, and G. J. Mueller. Forthcoming. Sovereign risk, fiscal policy, and macroeconomic stability. *Economic Journal.*
Davig, T., and E.M. Leeper. 2009. *Monetary–fiscal policy interactions and fiscal stimulus*. NBER Working Paper No. 15133 (July).
DeLong, B.J. 2010. Helicopter drop time: Paul Krugman gets one wrong. *Grasping Reality with Both Invisible Hands: Fair, Balanced, and Reality-Based: A Semi-Daily Journal* (14 July). Available at: http://delong.typepad.com/sdj/2010/07/helicopter-drop-time-paul-krugman-gets-one-wrong.html. Accessed 15 Feb 2013.
DeLong, B.J., and L. Summers. 2012. Fiscal policy in a depressed economy. *Brookings Papers on Economic Activity* 1: 233.
Eggertsson, G.B., and P. Krugman. 2012. Debt, deleveraging, and the liquidity trap: A Fisher–Minsky–Koo approach. *Quarterly Journal of Economics* 127(3): 1469–1513.
Ericsson, N., and J. Irons. 1995. The Lucas critique in practice: Theory without measurement. In *Macroeconometrics: Developments, tensions, and prospects*, ed. K.D. Hoover. New York: Springer.
Fazzari, S.M., J. Morley, and I. Panovska. 2012. *State dependent effects of fiscal policy.* Australian School of Business Research Paper No. 2012 ECON 27.
Fernández-Villaverde, J. 2010. Fiscal policy in a model with financial frictions. *American Economic Review* 100(2): 35–40.
Gali, J., J.D. López-Salido, and J. Vallés. 2007. Understanding the effects of government spending on consumption. *Journal of the European Economic Association* 5(1): 227–270.
Giavazzi, F., and M. Pagano. 1990. Can severe fiscal consolidations be expansionary? Tales of two small European countries. *NBER Macroeconomics Annual* 5: 75–111.
Goldberger, A.S. 1959. *Impact multipliers and dynamic properties of the Klein–Goldberger model*. Amsterdam: North-Holland Publishing Company.
Ilzetzki, E., E.G. Mendoza, and C.A. Vegh. Forthcoming. How big (small?) are fiscal multipliers?. *Journal of Monetary Economics* 60: 239.
IMF. 2010. Chapter 3: Will it hurt? Macroeconomic effects of fiscal consolidation. *World Economic Outlook* (October).
Kumhof, M., D. Laxton, D. Muir, and S. Mursula. 2010. *The Global Integrated Monetary and Fiscal Model (GIMF) – Theoretical structure*, IMF Working Paper No. 10/34. Washington, DC: International Monetary Fund.
Laxton, D., P. Isard, H. Faruqee, E. Prasad, and B. Turtelboom. 1998. *MULTIMOD Mark III: The core dynamic and steady-state models*, Occasional Paper 164. Washington, DC: IMF.
Mendel, B. 2012. Local multipliers: Theory and evidence. *Mimeo.* Harvard University, September.
Moretti, E. 2010. Local multipliers. *American Economic Review: Papers & Proceedings* 100: 1–7, May.
Mountford, A., and H. Uhlig. 2009. What are the effects of fiscal policy shocks? *Journal of Applied Econometrics* 24(6): 960–992.
Nakamura, E., and J. Steinsson. 2011. *Fiscal stimulus in a monetary union: Evidence from U.S. regions*, NBER Working Paper No. 17391. Cambridge, MA: National Bureau of Economic Research.
Perotti, R. 2011. The 'Austerity Myth': Gain Without Pain? *NBER Working Paper* No. 17571.
Ramey, V.A. 2011a. Identifying government spending shocks: It's all in the timing. *Quarterly Journal of Economics* 126(1): 1–50.
Ramey, V.A. 2011b. Can government purchases stimulate the economy? *Journal of Economic Literature* 49(3): 673–685.
Reichling, F., and C. Whalen. 2012. *Assessing the short-term effects on output of changes in federal fiscal policies*, Working Paper No. 2012–08. Washington, DC: Congressional Budget Office.
Romer, C.D. 2011. What do we know about the effects of fiscal policy? Separating evidence from ideology. Speech at Hamilton College, 7 November.
Romer, C.D., and D.H. Romer. 2010. The macroeconomic effects of tax changes: Estimates based on a new measure of fiscal shocks. *American Economic Review* 100: 763–801.
Romer, C. D. and D.H. Romer. 1989. Does monetary policy matter? A new test in the spirit of Friedman and Schwartz, *NBER Macroeconomics Annual 1989.* Cambridge, MA: MIT Press.
Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48(1): 1–48.
Spilimbergo, A., S. Symansky, and M. Schindler. 2009. *Fiscal multipliers*, IMF Staff Position Notes No. 09/11. Washington, DC: IMF.
Taylor, J.B. 1993. *Macroeconomic policy in a world economy*. New York: W.W. Norton.
Van Brusselen, P. 2009. *Fiscal stabilisation plans and the outlook for the world economy*, NIME Policy Brief No. 01–2009. Brussels: Federal Planning Bureau (April).
Woodford, M. 2011. Simple analytics of the government expenditure multiplier. *American Economic Journal: Macroeconomics* 3(1): 1–35.

Fiscal Stance

Terry Ward

Fiscal stance is commonly understood to denote the expansionary or contractionary implications for the economy of a government's budgetary policy. More precisely, it represents an attempt to summarize, in a single measure, the combined effect on aggregate demand, and therefore potentially on real output and income, of all the various decisions taken by government in respect of public expenditure, taxation and other sources of revenue which go to make up a national budget. As such it presupposes not only that governments can affect demand in this way, but also that it is possible to devise an indicator of this kind which is sufficiently widely accepted as to be useful. This later proposition has been questioned even by a number of self-professed Keynesian economists, who have emphasized the difficulty both of aggregating the effects of the many different items included in the budget and of disentangling these from other potential influences on demand.

The most straightforward way of producing an indicator of budgetary policy is to sum the inflows of revenue and outflows of expenditure to which they give rise and to take the difference between the two, the budget balance, as a measure of fiscal stance. Indeed any assessment of a government's macroeconomic policy commonly tends to focus on this magnitude. The difficulty with this figure, as was recognized almost as soon as proposals were first made to use the budget as a tool of economic management, is that 'it fails to distinguish the budget's influence on the economy from the economy's influence on the budget' (Okun and Teeters 1970, pp. 77–8). In other words, it incorporates both the consequences for budgetary flows of the tax rates set by government and the public expenditure outlays authorized by it from the effects on such flows of changes in income and expenditure in the economy. If there is a downturn in economic activity and income and expenditure grow less rapidly than usual, or even contract, then the revenue produced by a given set of tax rates will be correspondingly depressed and public expenditure will tend to be pushed up insofar as unemployment increases and the financial position of state (or publically supported) enterprises deteriorates (and *vice versa* if there is an upturn in activity). The problem is to separate these autonomous consequences from the discretionary effects of policy and thereby to distinguish the injection or withdrawal of purchasing power emanating from policy decisions from other sources of demand expansion or contraction, such as private sector borrowing or net export growth.

The solution, first devised in the 1940s, is to calculate budgetary flows at full employment levels of income and expenditure and to take the budget balance which would have resulted had GDP (or GNP) continuously followed such a growth path as the measures of fiscal stance. Indeed the term 'fiscal stance' has become synonomous with figures for the budget balance adjusted or normalized in this way. (The first estimate of such an adjusted balance seems to have been made by Kaldor 1944, for the UK for the year 1938. The concept was first proposed in the USA by the Committee of Economic Development in 1947, though the most influential was probably Brown 1956. On the US origins of the concept, see Blinder and Solow 1974.)

Such calculations served a dual purpose. They were used not only to indicate the expansionary or contractionary nature of budgetary policy, but also to reveal whether and to what extent the current stance of policy could be sustained in the longer term as full employment was approached. Accordingly, in a number of countries, the United States, West Germany and the Netherlands in particular, the 'full employment surplus' or 'structural budget balance' became widely accepted as a useful benchmark for assessing policy. (In the USA, it was included in the Annual Report of the Council of Economic Advisors and in the Reports of the Joint Economic Committee of Congress as well as in the Budget documents. In the Netherlands, it was included in the Budget Memorandum – see Netherlands Ministry of

Finance 1970 and Budget, 1978. In Germany, it was also used to assess policy as described in Dernberg 1975; and Chand 1978. For a discussion of alternative measures, see Lotz 1971.)

As time went on, as full employment in most countries became more remote, the full employment budget concept became less meaningful as a benchmark for policy, increasingly directed at objectives other than managing demand to secure particular rates of economic growth. By then, moreover, criticism of the concept as an indicator of fiscal stance was already widespread. Among the most frequently voiced concerns were: that the measure was not independent of the level of economic activity taken as the basis for normalization; that it was affected by the composition of the budget as well as by the difference between expenditure and revenue flows; that it made no allowance for the effect of inflation; that it ignored how the budget deficit was financed and more generally what kind of monetary policy was being followed; and, more recently, that it took not account of expectations and their influence on the effect of policy on the economy.

All of these criticisms are valid in some degree. The key issue, however, concerns the degree of validity and how far it is possible to modify the measure of fiscal stance to take account of them, without making it so complicated and so model-dependent that it ceases to be widely accepted as a satisfactory indicator of policy.

Thus while the level of economic activity chosen as the benchmark for standardizing the budget clearly affects the absolute value of the figures calculated, it tends to have much less effect on changes in the balance over time (see Ward and Neild 1978, pp. 33–7). Since fiscal stance can only be interpreted meaningfully in a comparative sense, in relation to policy in different periods, it is movements in the balance which are the relevant consideration. Nevertheless the composition of domestic income and expenditure, and therefore potentially the tax and public spending flows generated, does tend to vary as activity changes. This source of difficulty, however, can readily be minimized by choosing a benchmark level of activity which is not too different from the actual level – or even, to go one step further, by changing the benchmark each year to coincide with the actual level, a picture of the changing fiscal stance being built up by cumulating successive year-to-year movements. A further problem is that there may be disagreements about the rate of growth required to ensure that activity remains constant. These disagreements, however, are not usually so great as to give rise to marked divergences in estimates of fiscal stance, except when calculated over a number of years at a time. In this case, all that is possible is to produce a range of estimates, with the range of growth rates on which they are based made explicit.

A potentially more serious problem arises from the likelihood that different components of the budget have different effects on demand, so making simple aggregation of the revenue and expenditure flows involved inappropriate and possibly misleading. This has led many (Blinder and Solow 1974, among others) to propose that the budget components should be weighted according to the extent to which they feed into consumption or investment expenditure rather than into savings or, at one stage removed, imports (so explicitly allowing, *inter alia*, for the possibility of a balanced budget multiplier). The difficulty with such proposals is not only the increased complexity of the calculation and the greater scope for disagreement over any which is produced, but also their focus on the initial demand effects rather than on the longer-term consequences. Thus it is clearly unrealistic to suppose that 'first round' leakages from the circular process of income and expenditure determination are in some way lost for ever. Variations in income stemming from budgetary changes may not immediately feed into spending, but to major extent they ultimately will do so unless there is a permanent change in the desire of individuals and companies to increase or reduce their net holdings of financial assets. Over the long term, therefore, there tends to be a relatively stable relationship between the private sector financial balance, or savings less investment, which is the relevant concept in this context, and private sector income. For different forms of expenditure and taxation, the speed at which spending responds may well vary but there may be little significant difference in the long-term effect.

This means that any measure of fiscal stance has to specify the period over which the effect of policy

is being estimated. The shorter the period, the more important are differential leakages as between budget items likely to be, the more do underlying economic circumstances come into play and the more complicated and uncertain is the process of estimation. Indeed without a fully fledged macroeconomic model with built-in behavioural functions and sufficient disaggregation of taxation and public expenditure, it is hard to see how any satisfactory estimate of short-term budgetary effects could be constructed. Such an estimate, however, is really a measure of fiscal impact rather than of fiscal stance. (Estimates of a demand-weighted measure of fiscal impact are, for example, regularly published by the UK National Institute of Economic and Social Research in its quarterly review.) The longer the period, the less important does weighting become. If the concern is to assess the cumulative effects of policy over a time horizon of a year or two, then differential leakages into net holdings of financial assets, i.e. savings less investment, ought not to be a significant problem for most items and an unweighted measure is unlikely to give misleading results.

Nevertheless, there are certain budgetary items, though usually relatively minor in scale, for which even the long-term effect on demand is likely to be small. These are lending, asset sales or purchases and other purely financial transactions which are included in total public sector borrowing (in the US partly in the Credit Budget) but not the public sector financial balance (though this is typically not true of purchases less sales of land and existing buildings) and which tend to affect income and wealth only marginally. The most sensible and straightforward course of action is to exclude these from a measure of fiscal stance.

On the other hand, differential leakages into imports may be more of a problem. There is usually a general tendency for public expenditure to involve a lower import content than private spending, at least at the first round, and a measure of fiscal stance not adjusted for this might therefore misrepresent the scale of long-term demand effects if policy is heavily concentrated on, say, expanding public expenditure or reducing taxes.

The argument for adjusting measures of fiscal stance for inflation has two aspects. The first, and least serious, is that variations in inflation can affect tax revenue differently from public expenditure outlays, insofar as the two sides of the account are indexed to differing degrees. In most advanced economies, government revenue tends to increase in proportion to nominal income, or more than in proportion where the tax structure is progressive, while public expenditure sometimes lags behind inflation because of spending authorizations, or budget allocations, being specified in cash terms. In such circumstances, the budget deficit would be reduced if inflation were to increase without any overt action on the part of government. Though perhaps not intentional, a change of this kind ought to be treated as a tightening of fiscal stance in the same way as a deliberate increase in nominal tax rates which produced the same effect. To do otherwise would be to confuse action with intent and to regard inaction as signifying no change in policy even though it might be associated with significant changes in *effective* rates of taxation. This is accomplished by taking nominal changes in revenue and public expenditure in relation to nominal national income both measured in terms of actual prices, as the appropriate basis for calculating fiscal stance. Any change in this measure can then be interpreted as indicating a discretionary change in budgetary policy irrespective of its origins.

The more substantive aspect is that inflation can affect the real value of government debt in the economy, and therefore the real wealth of holders of government securities and presumably in turn their expenditure behaviour (see Tobin and Buiter 1976; Taylor and Threadgold 1979; Tanzi 1984). If, for example, such holders are not compensated for the erosion in their wealth caused by an increase in inflation relative to interest rates on the debt, then expenditure will tend to be depressed insofar as it is a function of wealth as well as current income. Conversely if interest rates lag behind prices when inflation falls, then debt holders will enjoy an increase in their real wealth which may tend to boost demand, in the longer term if not immediately. To ignore these effects is liable to give a misleading indication of fiscal stance. The cyclically adjusted budget balance ought, therefore, to be further adjusted to allow for the impact

of inflation on the real value of outstanding government debt (see Price and Muller 1984; OECD 1984). The difficulty is that the inflation rate which is relevant in this context is the expected future rate rather than the present rate. Since the former is unknown, there seems little practical alternative but to use the latter even though it is less than satisfactory (but see, e.g., Buiter 1983).

The expansionary effect of higher real interest rates on spending by holders of government debt is liable to be offset by a depressing effect on investment and consumption of durable goods from the higher cost of borrowing. The question arises as to how far this and other financial effects resulting from the monetary policy being followed by the government at the time should be taken into account in the measurement of fiscal stance. In principle, it can be argued that fiscal and monetary policy should be kept separate and the effects of the two on the economy estimated individually. In practice it is not quite so simple. Even though governments have some discretion over how to finance a budget deficit – whether by expanding the money supply or selling public sector debt to the non-bank private sector and abroad – the reaction of financial markets cannot simply be ignored and in reality the two strands of policy will be considered together.

Moreover, the possible financial and wider consequences of fiscal action, both internally and externally, might themselves affect the way that demand responds to such action. For example, in a world of floating exchange rates, the exchange value of a country's currency might itself be partly determined by the fiscal policy being followed, so that a larger budget deficit might lead to a fall in the exchange rate (or possibly a rise if interest rates are expected to go up) and a stimulus to demand from net exports as well as from fiscal policy directly. Alternatively, anticipations about the way a government might respond to prevent such a fall through modifying its monetary policy (by raising interest rates for example, or tightening credit) might itself influence the speed and scale of the internal demand response to the fiscal measures introduced.

More generally, expectations about future developments and the close relationship between the budget and other aspects of policy and other sources of demand generation represent potentially serious problems for measuring fiscal stance in any simple, straightforward manner. Thus in addition to any effects on interest rates and exchange rates, a decision to increase the budget deficit – in the present, for example – might be taken to imply a need for higher taxes in the future to service the additional debt created and hence might generate little increase in demand, to the extent that expenditure is determined by expected income over the long-run rather than current income. The argument, in its extreme version, is that reactions to the expected consequences for future public expenditure, taxation and the budget balance of present budgetary decisions are liable to frustrate the expansionary or contractionary intentions of government more or less completely. In a highly uncertain world, however, it is hardly plausible that such anticipations would fully offset attempts by government to manage demand, though it is not implausible that they might modify the effects of policy in some degree. Nor is it implausible that the degree of influence might vary according to what else is happening in the economy at the time.

In view of these considerations, it is futile to hope that any simple, easily constructed measure of fiscal stance is likely to capture fully the effects of budgetary policy at all moments in time. The question which remains is whether the only resort is to macroeconomic models which are sufficiently detailed and reliable to enable the effects of any particular package of fiscal measures to be isolated from other influences on demand (as advocated, for example, by Buiter 1985). But in this case, the purpose of such an exercise would be unclear since the main concern is presumably with the combined effect of government policy taken as a whole rather than with any individual part of it.

In reality it is hard to believe that a measure of fiscal stance adjusted for inflation and cyclical variations in economic activity has no useful role to play in assessing government policy, despite its drawbacks and despite the heavy qualifications which ought to surround its use. Certainly the regular publication of such measures by the IMF and OECD seems to make a valuable contribution to the policy debate. At the very least, it provides an important counterbalance to the focus on the

actual budget deficit which has been a feature of policy discussion in most countries in the 1970s and 1980s, which in many cases has led to the stance of policy being seriously misrepresented and which has therefore contributed to perverse policy action being taken.

See Also

- Budgetary Policy
- Full Employment Budget Surplus
- Functional Finance
- Stabilization Policy

References

Blinder, A.S., and R.M. Solow. 1974. Analytical foundations of fiscal policy. In *The economics of public finance*, ed. A.S. Blinder et al. Washington, DC: Brookings Institution.

Brown, E.C. 1956. Fiscal policy in the thirties: A reappraisal. *American Economic Review* 46: 857–879.

Buiter, W.H. 1983. The theory of optimum deficits and debt. In *The economics of large government deficits*. Boston: Federal Reserve Bank of Boston.

Buiter, W.H. 1985. A guide to public sector debt and deficits. *Economic Policy* 1: 13–79.

Chand, S.K. 1978. Summary measures of fiscal influence. In *IMF Staff Papers* 24, Washington, DC.

Committee for Economic Development. 1947. *Taxes and the budget: A program for prosperity in a free economy*. Washington, DC.

Dernberg, T.F. 1975. Fiscal analysis in the Federal Republic of Germany: The cyclically neutral budget. *IMF Staff Papers* 22: 825–827.

Kaldor, N. 1944. Appendix C of W.H. Beveridge. In *Full employment in a free society*. London: Allen & Unwin.

Lotz, J. 1971. *Techniques of measuring the effects of fiscal policy*, OECD economic outlook, Occasional studies. Paris: OECD.

Netherlands Ministry of Finance. 1970. *The Netherlands budget memorandum 1970*, Annex 2. The Hague.

OECD. 1984. *Economic outlook*. Paris: OECD.

Okun, A.M., and Teeters, N.H. 1970. The full employment surplus revisited. *Brookings Papers on Economics Activity* No. 1, 770–110.

Price, R.W.R., and Muller, P. 1984. Structural budget indicators and the interpretation of fiscal policy stance in OECD 'economies'. *OECD Economic Studies* (3): 27–72.

Tanzi, V. (ed.). 1984. *Taxation, inflation and interest rates*. Washington, DC: IMF.

Taylor, C.T., and Threadgold, A.R. 1979. *Real national savings and its sectoral composition. Bank of England Discussion Papers* No. 6. London: Economic Intelligence Department, Bank of England.

Tobin, J., and W.H. Buiter. 1976. Long run effects of fiscal and monetary policy on aggregate demand. In *Monetarism*, ed. J.L. Stein. Amsterdam: North-Holland.

Ward, T.S., and R.R. Neild. 1978. *The measurement and reform of budgetary policy*. London: Heinemann.

Fiscal Theory of the Price Level

Marco Bassetto

Abstract

The fiscal theory of the price level (FTPL) describes fiscal and monetary policy rules such that the price level is determined by government debt and fiscal policy alone, with monetary policy playing at best an indirect role. This theory clashes with the monetarist view that states that money supply is the primary determinant of the price level and inflation. Furthermore, many authors have argued that the fiscal rules upon which the FTPL relies are misspecified. We review the sources of disagreement, and highlight aspects upon which some consensus has emerged.

Keywords

Budget deficits; Commodity money; Debt crises; Dynamic competitive equilibrium; Exogenous interest rates; Fiscal theory of the price level; Government budget constraint; Inflation; Interest rate peg; Interest rate rules; Intertemporal budget constraint; Monetarism; Monetary policy; Monetization of government debt; Money supply rule; Multiple equilibria; Nominal interest rate; Price level; Quantity theory of money; Real money balances; Seigniorage; Sunspots; Uniqueness of equilibrium; Velocity of circulation

JEL Classifications

D4; D10

The fiscal theory of the price level (FTPL) describes policy rules such that the price level is determined by government debt and the present and future tax and spending plans, with no direct reference to monetary policy.

In understanding the FTPL and tracing its roots, we start from two simple relations: the velocity equation and the government budget constraint.

The velocity equation defines the velocity of money in period t (V_t) as the ratio of nominal output (the price level P_t times real output Y_t) to nominal money balances (M_t):

$$V_t = \frac{P_t Y_t}{M_t}, t = 0, 1, \ldots \quad (1)$$

Differences across monetary models arise in the way these four economic variables are determined, and in the specification of which (if any) of these variables is to be treated as exogenous as opposed to endogenous. Prior to the introduction of the FTPL, Eq. (1) was viewed as the primary determinant of the price level. As an example, the quantity theory of money states that V_t is fixed and exogenous. In this case, the price level is proportional to the money supply. High prices arise because too much money is chasing too few goods, which is the heart of the monetarist doctrine. In a more sophisticated theory, velocity is itself affected by other macroeconomic variables, chief among them the nominal interest rate. Furthermore, in general, the price level needs to be determined jointly with M_t, Y_t, and V_t by computing the entire equilibrium path of the economy. The FTPL traces its roots to an incompleteness in the monetarist view of the price level: often, the equilibrium price level fails to be uniquely determined, that is, there are many paths of P_t that satisfy (1) as well as all the other equilibrium requirements (see discussion in Kocherlakota and Phelan 1999). This is especially true when monetary policy prescribes an exogenous interest rate; Sargent and Wallace (1975) show that the initial price level is then indeterminate, and subsequent inflation is subject to 'sunspots', uncertainty driven by self-fulfilling expectations. In the simplest case, an interest-rate peg determines the level of velocity (V_t), and real output and interest rates are independent of money and prices; eq. (1) then pins down *real* money balances (M_t/P_t), but it does not specify whether those balances will be attained by high or low nominal money supply and prices.

The FTPL (Woodford 1994; Sims 1994) determines prices from a different equation:

$$\frac{B_t}{P_t} = \text{Present value of primary surpluses as of fiscal time } t, t = 0, 1, \ldots, \quad (2)$$

where Bt is the nominal value of government liabilities (debt and money) at the beginning of period t. Equation (2) is the government budget constraint, in its present value form: the left-hand side represents real government liabilities, matched by assets on the right-hand side. In its simplest form, the FTPL assumes that the government commits to a fixed and exogenous present value of primary fiscal surpluses; this is a special case of what Leeper (1991) defines as an 'active' fiscal policy and Woodford (1995) a 'non-Ricardian' fiscal regime. Given an initial condition for debt, B_0, a unique price level is consistent with (2): the FTPL successfully selects a unique price level at time 0, even in the case of an interest rate peg, for which the monetarist view offered no prediction. The power of the FTPL is not limited to period 0; the possibility of sunspot equilibria is ruled out in all subsequent periods, since again a unique level of prices is consistent with a given present value of surpluses and the nominal debt inherited from the past. Nonetheless, monetary policy does have an effect on inflation after period 0: the evolution of nominal liabilities B_t depends on the nominal interest rate, which is affected by monetary policy.

Since its inception, the FTPL has been extremely controversial. I focus here on two main areas of concern.

The Value of Money or the Value of Debt?

The price level is defined as the inverse of the value of money: how much money it takes to buy

a given basket of goods. By contrast, the FTPL is about the inverse of the value of *government debt*. This is explained particularly clearly in Cochrane (2005). As Buiter (2002) points out, there is no reason in general for the value of debt and the value of money to coincide. To the extent that households anticipate a government default, they may trade government debt at a discount, without necessarily affecting the value of money. This criticism is particularly serious when the central bank adopts a monetary policy that rules out monetization of government debt. As an example, consider the case in which the central bank adopts a constant money supply rule and does not engage in open market operations. In this case, there is no link between government debt and money, and no reason why a maturing T-Bill with a face value of $1,000 should trade at par with ten $100 notes issued by the central bank. Maturing debt and money will trade at par only if *fiscal policy* is run in such a way that the government will have the appropriate amounts of money to repay its debt, independently of the price level: this requires real tax revenues to adjust to prices, violating the central assumption of the FTPL.

The same criticism does not apply when the monetary policy of the central bank allows unlimited monetization of debt, as in the case of an interest rate peg. In this case, the central bank commits to exchange arbitrary amounts of money and one-period government debt at a fixed price. This commitment is not inconsistent with a second commitment, to redeem all maturing government debt at par, offering money in exchange. Since the central bank has unlimited ability to produce money, a government default on nominal debt is now ruled out. In this case, the FTPL is simply a version of a commodity money standard; money, as well as other government liabilities, is backed by the present value of future government surpluses, just as the value of Microsoft shares is backed by the present value of Microsoft profits (this is the main example in Cochrane 2005).

While the original treatment of the FTPL was ambiguous (in particular, Woodford 1995, considers the case of the FTPL under a money supply rule), it is now widely agreed that the FTPL requires an implicit or explicit institutional commitment to prevent a government default (or excess repayments by the government) through an appropriate (de)monetization of debt. In this form, the FTPL bears some similarities with the 'unpleasant monetarist arithmetic' of Sargent and Wallace (1981). Under the monetarist arithmetic, a fiscal deficit imbalance will trigger inflation, because seigniorage revenues are necessary to prevent the government from defaulting. Even though monetization of government debt plays a central role in both theories, there are important differences. According to the unpleasant monetarist arithmetic, seigniorage revenues (which are part of the present value of surpluses in Eq. (2)) will have to respond to changes in Pt to ensure that the government budget constraint holds; hence, equilibrium occurs through adjustments in the right-hand side of (2). In the FTPL, seigniorage revenues on the monetary base play at best a minor role. Under the FTPL, it is the price level that responds to shocks to spending and taxes; its fluctuations cause the real value of debt (the left-hand side of (2)) to appreciate or depreciate to reach an equilibrium.

Government Constraints and Equilibrium Conditions

The FTPL is based on the assumption that Eq. (2) holds only at an equilibrium. The critics of the FTPL view instead (2) as a constraint that forces the government to match the real value of debt with an appropriate present value of primary surpluses, for all conceivable levels of prices. To better understand the issue, it is useful to note that (2) looks identical to the intertemporal budget constraint of any household in the economy: it is sufficient to relabel Bt as the nominal liabilities of the household, and the right-hand side as the present value of its non-asset income, net of consumption. In the case of a household, there is universal agreement that (2) should be viewed as a constraint: given any value of Pt, the household must choose a consumption/income plan that satisfies (2). The critics of the FTPL argue that the government should be no different from any other

agent. Unlike the previous criticism, the heated debate that has emerged on this point has not resulted in widespread agreement. As Bassetto (2005) points out, the disagreement stems from a fundamental weakness in the tools that have been used to study this problem. Both critics and supporters of the FTPL adopt the dynamic competitive equilibrium framework. This framework is designed for environments populated by many small players; in the presence of a large and potentially strategic player, such as the government, it offers little guidance in distinguishing between equilibrium conditions and constraints that the large player faces under any circumstances, even away from an equilibrium. While there are many applications for which this ambiguity is not important, a proper account of the distinction is essential to study the uniqueness or multiplicity of equilibria, which is the object of interest in the case of the FTPL.

To overcome this difficulty, Bassetto (2002) explicitly describes the economy as a game, where the actions available to all households and the government at any point in time are clearly spelt out. Bassetto shows that the basic version of the FTPL, with an unconditional commitment to a sequence of primary surpluses, is not a valid government strategy in a well-specified game, at least if the sequence includes a primary deficit at any point in time. Intuitively, a primary deficit is possible only if the government is able to raise revenues through borrowing. Since lending is voluntary (unlike payment of taxes), any plausible game includes the possibility that private agents will not lend; if this circumstance arises, the government is forced to a fiscal adjustment. Bassetto then proves that there exist other government strategies that lead to a unique equilibrium price level that is determined from taxes and spending alone. These strategies paint a very different picture of the conditions under which a FTPL arises: whereas the traditional view relies on the government setting taxes and spending exogenously, with no regard for the evolution of debt, the strategies described by Bassetto require the government to strongly react to incipient 'debt crises' by accumulating larger surpluses in present value.

Empirical Studies

A small empirical literature (for example, Canzoneri et al. 2001; Cochrane 2001) has looked into the usefulness of (2) in accounting for the evolution of prices. The results are not very favourable; in particular, when a government runs an unexpected deficit, the real market price of its debt increases, suggesting that households expect that the government will make up for the shortfall through increased surpluses in the future. If future surpluses were exogenous and fixed, (2) would suggest that an unexpected deficit should have its primary effect through inflation, by depressing the real market value of debt. While these observations cannot refute the central claim of the FTPL, that (2) is only an equilibrium condition, they call into question the usefulness of the FTPL in explaining actual inflationary episodes.

Conclusion

Recent research into monetary policy has looked for interest rate rules that ensure price level determinacy independently of the fiscal policy of the government; this has weakened interest in the FTPL. Though no issue as controversial as the FTPL has emerged since, this recent analysis is still open to ambiguous distinctions between policy rules, which should capture government behaviour in all possible scenarios, and equilibrium relations across the endogenous variables of an economic system. A more complete analysis awaits the development of new tools that are as simple and powerful as dynamic competitive equilibrium, and yet able to appropriately capture the special role of the government.

See Also

- ▶ Commodity Money
- ▶ Determinacy and Indeterminacy of Equilibria
- ▶ Government Budget Constraint
- ▶ Monetarism
- ▶ Monetary and Fiscal Policy Overview
- ▶ Multiple Equilibria in Macroeconomics

- ▶ Quantity Theory of Money
- ▶ Sunspot Equilibrium

Bibliography

Bassetto, M. 2002. A game-theoretic view of the fiscal theory of the price level. *Econometrica* 70: 2167–2195.
Bassetto, M. 2005. Equilibrium and government commitment. *Journal of Economic Theory* 124: 79–105.
Buiter, W. 2002. The fiscal theory of the price level: A critique. *Economic Journal* 112: 459–480.
Canzoneri, M., R. Cumby, and B. Diba. 2001. Is the price level determined by the needs of fiscal solvency? *American Economic Review* 91: 1221–1238.
Cochrane, J. 2001. Long term debt and optimal policy in the fiscal theory of the price level. *Econometrica* 69: 69–116.
Cochrane, J. 2005. Money as stock. *Journal of Monetary Economics* 52: 501–528.
Kocherlakota, N., and C. Phelan. 1999. Explaining the fiscal theory of the price level. *Federal Reserve Bank of Minneapolis Quarterly Review* 23(4): 14–23.
Leeper, E. 1991. Equilibria under 'active' and 'passive' monetary policies. *Journal of Monetary Economics* 27: 129–147.
Sargent, T., and N. Wallace. 1975. 'Rational' expectations, the optimal monetary instrument, and the optimal money supply rule. *Journal of Political Economy* 83: 241–254.
Sargent, T., and N. Wallace. 1981. Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis Quarterly Review* 5(3): 1–17.
Sims, C. 1994. A simple model for study of the determination of the price level and the interaction of monetary and fiscal policy. *Economic Theory* 4: 381–399.
Woodford, M. 1994. Monetary policy and price level determinacy in a cash-in-advance economy. *Economic Theory* 4: 345–380.
Woodford, M. 1995. Price level determinacy without control of a monetary aggregate. *Carnegie-Rochester Conference Series on Public Policy* 43: 1–46.

Fisher, Irving (1867–1947)

James Tobin

JEL Classifications
B31

Irving Fisher was born in Saugerties, New York, on 27 February 1867; he was residing in New Haven, Connecticut at the time of his death in a New York City hospital on 29 April 1947.

Fisher is widely regarded as the greatest economist America has produced. A prolific, versatile and creative scholar, he made seminal and durable contributions across a broad spectrum of economic science. Although several earlier Americans, notably Simon Newcomb, had used some mathematics in their writings. Fisher's dedication to the method and his skill in using it justify calling him America's first mathematical economist. He put his early training in mathematics and physics to work in his doctoral dissertation on the theory of general equilibrium. Throughout his career his example and his teachings advanced the application of quantitative method not only in economic theory but also in statistical inquiry. He, together with Ragnar Frisch and Charles F. Roos, founded the Econometric Society in 1930; and Fisher was its first President. He had been President of the American Economic Association in 1918.

Much of standard neoclassical theory today is Fisherian in origin, style, spirit and substance. In particular, most modern models of capital and interest are essentially variations on Fisher's theme, the conjunction of intertemporal choices and opportunities. Likewise, his theory of money and prices is the foundation for much of contemporary monetary economics.

Fisher also developed methodologies of quantitative empirical research. He was the greatest expert of all time on index numbers, on their theoretical and statistical properties and on their use in many countries throughout history. From 1923 to 1936, his own Index Number Institute manufactured and published price indexes of many kinds from data painstakingly collected from all over the world. Indefatigable and innovative in empirical research, Fisher was an early and regular user of correlations, regressions and other statistical and econometric tools that later became routine.

To this day Fisher's successors are often rediscovering, consciously or unconsciously, Fisher's ideas and building upon them. He can be credited with distributed lag regression, life cycle saving theory, the 'Phillips curve', the case

for taxing consumption rather than 'income', the modern quantity theory of money, the distinction between real and nominal interest rates, and many more standard tools in economists' kits. Although Fisher was not fully appreciated by his contemporaries, today he leads other old-timers by wide and increasing margins in journal citations. In column inches in the *Social Sciences Citation Index* (1979, 1983), Fisher led his most famous contemporaries, Wesley Mitchell, J.B. Clark, and F.W. Taussig in that order, by rough ratios 5:3:1:1 in 1971–5 and 9:3:1:1 in 1976–80. Much more than the others, moreover, Fisher is cited for substance rather than for history of thought.

For all his scientific prowess and achievement, Fisher was by no means an 'ivory tower' scholar detached from the problems and policy issues of his times. He was a congenital reformer, an inveterate crusader. He was so aggressive and persistent, and so sure he was right, that many of his contemporaries regarded him as a 'crank' and discounted his scientific work accordingly. Science and reform were indeed often combined in Fisher's work. His economic findings, theoretical and empirical, would suggest to him how to better the world; or dissatisfaction with the state of the world would lead him into scientifically fruitful analysis and research. Fisher's search for conceptual clarity about 'the nature of capital and income' led him not only to lay the foundations of modern social accounting but also to argue that income taxation wrongly puts saving in double jeopardy. Fisher turned his talents to monetary theory because he suspected that economic instability was largely the fault of existing monetary institutions. His 'debt-deflation theory of depression' was motivated by the disasters the Great Depression visited upon the world.

Economics was not the only aspect of human and social life that engaged Fisher's reformist zeal. He was active and prolific in other causes: temperance and Prohibition; vegetarianism, fresh air, exercise and other aspects of personal hygiene; eugenics; and peace through international association of nations.

Fisher was an amazingly prolific and gifted writer. The bibliography compiled by his son lists some 2000 titles authored by Fisher, plus another 400 signed by his associates or written by others about him. Fisher's writings span all his interests and causes. They include scholarly books and papers, articles in popular media, textbooks, handbooks for students, tracts, pamphlets, speeches and letters to editors and statesmen. They include the weekly releases of index numbers, often supplemented by commentary on the economic outlook and policy, issued for thirteen years by Fisher and assistants from the Index Number Institute housed in his New Haven home.

Fisher was the consummate pedagogical expositor, always clear as crystal. He hardly ever wrote just for fellow experts. His mission was to educate and persuade the world. He took the trouble to lead the uninitiated through difficult material in easy stages. Whenever he was teaching or tutoring students, he wrote handbooks or texts for their benefit – in mathematics and science when he was still a student himself, in the principles of economics when he was the professor responsible for the introductory course. Fisher's economics text was published in 1910 and 1911. Its graceful exposition of sophisticated theoretical material will impress a modern connoisseur, but it was too difficult for widespread adoption. Some of it survived in a leading introductory text of the 1920s and 1930s, by the younger Yale economists Fairchild et al. (1926).

A Brief Biography

Irving Fisher grew up and attended school successively in Peace Dale, Rhode Island; New Haven, Connecticut; and St Louis, Missouri. His father, a Congregational minister, died of tuberculosis just when Irving had finished high school and was planning to attend Yale College, his father's *alma mater*. Irving was now the principal breadwinner for himself, his mother and his younger brother. He did have a $500 legacy from his father for his college education. The family moved to New Haven, and together managed to make ends meet. Irving tutored fellow students during term and in summers.

Fisher was a great success in Yale College, ranking first in his class and winning prizes and

distinctions not only in mathematics but across the board. He was also determined to make good in the extra-curricular college culture so important in those days. His efforts won him election to the most prestigious secret senior society, Skull and Bones, the ultimate reward senior campus leaders bestowed on members of the class behind them.

Awarded a scholarship for graduate study, he stayed on at Yale. Graduate Studies were not departmentalized in those days, and Fisher ranged over mathematics, science, social science and philosophy. His most important teachers were Josiah Willard Gibbs, the mathematical physicist celebrated for his theory of thermodynamics, William Graham Sumner, famous still in sociology but at the time also important in political economy, and Arthur Twining Hadley, a leading economist specializing in what is now known as Industrial Organization.

As the time to write a dissertation approached, Fisher had still not chosen his life work. Young Fisher's interests and talents were universal. In the seven years at Yale before he finished his doctorate, he had written and published poetry, political commentary, book reviews, a geometry text together with tables of logarithms, and voluminous notes on mathematics, mechanics and astronomy for the benefit of students he was teaching or tutoring. If he had specialized in anything in six years at Yale, it was mathematics, but even in his graduate years he had spent half his time elsewhere.

Summer put him on to mathematical economics, and in his third year of graduate study, he finished the dissertation that won him worldwide recognition in economic theory. Fisher's 1891 PhD was the first one in pure economics awarded by Yale, albeit by the faculty of mathematics. Although the university, thanks to Summer, Hadley and Henry W. Farnum, was strong in 'political economy', there was no distinct department for the subject, let alone for 'economics'. This was generally the case in American universities. Venturing into mathematical economic theory, Fisher was very much on his own; and his route into economics was quite different from that of most American economists of his era.

The dominant tradition in American political economy was imported from the English classical economists, mainly Smith, Ricardo and John Stuart Mill; it was just beginning to be updated by Marshall. This tradition Fisher's mentors at Yale had taught him well. But the neoclassical developments on the European continent from 1870 on, the works of Warlas and Menger and Böhm-Bawerk, or even those of their English counterparts Jevons and Edgeworth, had been little noticed at Yale or elsewhere in America.

At the time, the main challenge in America to classical political economy was coming from quite a different direction. The American Economic Association was founded in 1886 by young rebels against Ricardian dogma and its *laissez-faire* political and social message. They included Richard T. Ely, J.B. Clark, Edwin R.A. Seligman and other future luminaries of American economics. Many of them had pursued graduate studies in Germany. In the German emphasis on historical, institutional and empirical studies they found welcome relief from implacable classical theory, and in the German faith in the state as an instrument of socially beneficial reform they found a hopeful antidote to the fatalism of economic competition and social Darwinism. Sumner was prominent among several elders who refused to join an Association born of such heresy; he did not relent even though the AEA very soon became sufficiently neutral and catholic to attract his Yale colleagues and other initial holdouts. Fisher, a bit younger than the founding rebels and educated solely at one American university, was not involved. It was his reconstruction, rather than their revolution, that was destined eventually to replace the classical tradition in the mainstream of American economics.

Fisher stayed at Yale throughout his career. He started teaching mathematics, evidently even before he received his doctorate and was appointed Tutor in Mathematics. His first economics teaching was under the auspices of the mathematics faculty, an undergraduate course on 'The Mathematical Theory of Prices'. In 1894–5 during his *Wanderjahr* in Europe, this young American star was welcomed by the leading mathematically inclined theorists in every

country. On his return he became Assistant Professor of Political and Social Science and began teaching economics proper. He was appointed full Professor in 1898 and retired in 1935.

Fisher was struck by tuberculosis in 1898. He spent the first three years of his professorship on leave from Yale and from science, recuperating in more salubrious climates. His lifelong crusade for hygienic living dates from this personal struggle to regain health and vigour. The experience powerfully reinforced his determination to gain 'a place among those who have helped along my science' and his ambition 'to be a *great* man', as he wrote to his wife (I.N. Fisher 1956, pp. 87–8). After his recovery the books and articles began flowing from his pen, never to stop until his death at the age of 80.

Fisher participated actively in teaching and in university affairs until 1920. Thereafter his writings and his myriad outside activities and crusades preoccupied him. He taught only half time and had little impact on students, undergraduate or graduate. Thus Fisher had few personal disciples; there was no Fisherian School. The student to whom Fisher was closest, personally and intellectually, was James Harvey Rogers, a 1916 PhD who returned to Yale as a professor in 1930. His career was prematurely ended by his tragic death in a plane crash in 1939 at the age of 55.

Fisher was, on top of everything else, an inventor. His most successful and profitable invention was the visible card index system he patented in 1913. In 1925 Fisher's own firm, the Index Visible Company, merged with its principal competitor to form Kardex Rand Co., later Remington Rand, still later Sperry Rand. The merger made him wealthy. However, he subsequently lost a fortune his son estimated to amount to 8 or 10 million dollars, along with savings of his wife and her sister, when he borrowed money to exercise rights to buy additional Rand shares in the bull market of the late 1920s.

More than money was at risk in the market. Fisher had staked his public reputation as an economic pundit by his persistent optimism about the economy and stock prices, even after the 1929 crash. His reputation crashed too, especially among non-economists in New Haven, where the university had to buy his house and rent it to him to save him from eviction. Until the 1950s the name Irving Fisher was without honour in his own university. Except for economic theorists and econometricians, few members of the community appreciated the genius of a man who lived among them for 63 years.

Irving Fisher's marriage to Margaret Hazard in 1893 was a very happy one for 47 years. She died in 1940. They had two daughters and one son, his father's biographer. The death of their daughter Margaret in 1919 after a nervous breakdown was the greatest tragedy of her parents' lives. Their daughter Carol brought them two grandchildren.

General Equilibrium Theory

Fisher's doctoral dissertation (1892) is a masterly exposition of Walrasian general equilibrium theory. Fisher, who was meticulous about acknowledgements throughout his career, writes in the preface that he was unaware of Walras while writing the dissertation. His personal mentors in the literature of economics were Jevons (1871) and Auspitz and Lieben (1889).

Fisher's inventive ingenuity combined with his training under Gibbs to produce a remarkable hydraulic-mechanical analogue model of a general equilibrium system, replete with cisterns, valves, levers, balances and cams. Thus could he display physically how a shock to demand or supply in one of ten interrelated markets altered prices and quantities in all markets and changed the incomes and consumption bundles of the various consumers. The model is described in detail in the book; unfortunately both the original model and a second one constructed in 1925 have been lost to posterity. Anyway Fisher was a precursor of a current Yale professor, Herbert Scarf (1973) and other practitioners of computing general equilibrium solutions. In his formal mathematical model-building too, Fisher was greatly impressed by the analogies between the thermodynamics of his mentor Gibbs and economic systems, and he was able to apply Gibbs's innovations in vector calculus.

Fisher expounds thoroughly the mathematics of utility functions and their maximization, and he

is careful to allow for corner solutions. He uses independent and additive utilities of commodities in his first mathematical approximation and in his physical model; later he was to show how this assumption could be exploited to measure marginal utilities empirically (1927). But the general formulation in his dissertation makes the utility of every commodity depend on the quantities consumed of all commodities. At the same time, he states clearly that neither interpersonally comparable utility nor cardinal utility for each individual is necessary to the determination of equilibrium. Fisher's list of the limitations of his analysis is candid and complete. The supply side of Fisher's model is, as he acknowledges, primitive. Each commodity is produced at increasing marginal cost, but neither factor supplies and prices nor technologies are explicitly modelled.

Finally, Fisher shows his enthusiasm for his discovery of mathematical economics by appending to his dissertation as published an exhaustive survey and bibliography of applications of mathematical method to economics.

General Equilibrium with Intertemporal Choices and Opportunities

The distribution of income and wealth, and in particular the sources, determinants and social rationales, of interest and other returns to private property, were obsessive topics in economics, both in Europe and North America, at the turn of the century. One important reason, especially in Europe, was the Marxist challenge to the legitimacy of property income. Answering Marx was a strong motivation for the Austrian school, in particular for the capital theory of Böhm-Bawerk and his followers. Neoclassical economics was in a much better position than its classical precursor to respond to the Marxist challenge. The labour theory of value, which Marx borrowed from the great classical economists themselves, neither explains nor justifies functionally or ethically incomes other than wages.

These topics engaged the two leading American economists of the era, John Bates Clark and Fisher. Clark (1899) set forth his marginal productivity theory of distribution, arguing that a generalized factor of production, capital, the accumulation of past savings, has like labour a marginal product that explains and justifies the incomes of its owners.

Fisher attacked these problems in a more elegant, abstract, mathematical, general and ethically neutral manner than Clark, and than Böhm-Bawerk. At the same time, his approach was clearer, simpler and more insightful than that of Walras.

The general equilibrium system of Fisher's dissertation was a single-period model. No intertemporal choices entered; hence the theory was silent on the questions of capital and interest. But Fisher took up these subjects soon after.

His first contribution, one that should not be underestimated, was to set straight the concepts and the accounting. This he did in (1896) and (1906) with clarity and completeness that have scarcely been surpassed. It's all there: continuous and discrete compounding; nominal versus real rates; the distinction between high prices and rising prices, and its implications for observations of interest rates; the inevitable differences among rates computed in different *numéraires*; rates to different maturities and consistency among them; appreciation, expected and unexpected; present values of streams of in- and out-payments; and so on. Schumpeter calls this work 'the first economic theory of accounting' and says 'it is (or should be) the basis of modern income analysis' (1954, p. 872).

Perhaps the most remarkable feature is Fisher's insistence that 'income' is consumption, including of course consumption of the services of durable goods. In principle, he says, income is psychic, the subjective utility yielded by goods and services consumed. More practically, income could be measured as the money value, or value in some other *numéraire*, of the goods and services directly yielding utility, but only of those. Receipts saved and invested, for example in the purchase of new durable goods, are not 'income' for Fisher; they will yield consumption and utility later, and those yields will be income. To include both the initial investment and the later yields as income is, according to Fisher, as absurd as to

count both flour and bread in reckoning net output. This view naturally led Fisher to oppose conventional income taxation as double taxing of saving, and to favour consumption taxation instead. His views on these matters are loudly echoed today.

Fisher published his theory of the determination of interest rates in *The Rate of Interest* (1907). A revised and enlarged version was published in 1930 as *The Theory of Interest.* One motivation for the revision was that Fisher's many critics apparently did not understand the 1907 version. They typically concentrated on the 'impatience' side of Fisher's theory of intertemporal allocation and missed the 'opportunities' side. It was there in 1907 already; the theory is much the same in both versions.

In 1930 Fisher is at pains to label his theory the 'impatience and opportunity' theory. 'Every essential part of it', he acknowledges, 'was at least foreshadowed by John Rae in 1834.' He does claim originality for his concept of 'investment opportunity'. This turns on 'the rate of return over cost, [where] both cost and return are differences between two optional income streams' (1930, p. ix). As Keynes acknowledged, this is the same as his own 'marginal efficiency of capital' (Keynes 1936, p. 140).

In these books Fisher extended general equilibrium theory to intertemporal choices and relationships. This strategy was different from Walras. Walras tried to extend his multi-commodity multi-agent model of exchange to allow for production, saving and investment. This maintained his stance of full generality but was also difficult to expound and to understand. Fisher saw that intertemporal dependences were tricky enough to justify isolating them from the intercommodity complexities that had concerned him in his doctoral thesis. Therefore he proceeded as if there were just one aggregate commodity to be produced and consumed at different dates. This simplification enabled him to illuminate the subject more brightly than Walras himself.

The methodology of Fisher's capital theory is very modern. His clarifications of the concepts of capital and income lead him to formulate the problem as determination of the time paths of consumption – that is, income – both for individual agents and for the whole economy. Then he divides the problem into the two sides, tastes and technologies, that are second nature to theorists today. One need only read Böhm-Bawerk's murky mixture of the two in his list of reasons for the agio of future over present consumption to realize that Fisher's procedure was not instinctive in those times.

Fisher's theory of individual saving is basically the standard model to this day. Undergraduates learn the two-period 'Fisher diagram', where a family of indifference curves in the two commodities consumption now c_1 and consumption later c_2 confront a budget constraint $c_1 + c_2/(1 + r) = y_1 + y_2(1 + r)$, where the y's are exogenous wage incomes in the two periods and r is the (real) market interest rate. From the usual tangency can be read the consumption choices and present saving or dissaving. This is indeed a Fisher diagram, but of course he went much beyond it.

He stated clearly what we now call the 'life cycle' model, explaining why individuals will generally prefer to smooth their consumption over time, whatever the time path of their expected receipts. But he was not dogmatic, and he allowed room for bequests and for precautionary saving. Where Fisher differed from later theorists, and especially from contemporary model-builders, was in his unwillingness to impose any assumed uniformity on the preferences (or expectations of 'endowments' – the latter term was not familiar to him though the concept was) of the agents in his economies, and in his scruples against buying definite results by assuming tractable functional forms. In general, many of the advances claimed in present-day theory appear to depend on greater boldness in these respects.

On the side of technology, Fisher's approach was the natural symmetrical partner of his formulation of preferences, equally simple, abstract and general. He assumed that the 'investment opportunities' available to an individual (not necessarily the same for everybody) and to the society as a whole can be summarized in the terms on which consumption at any date can be traded, with 'nature', for consumptions at other dates. In modern language, we would say that Fisher postulated

intertemporal production possibility frontiers, properly convex in their arguments, consumptions at various dates.

All that remained for Fisher, then, was to assume complete intertemporal loan markets cleared by real interest rates, count equations, and show that in principle the equalities of saving and investment at every date determine all interest rates and the paths of consumption and production for all individuals and for the society. Like hundreds of mathematical theorists since, he set the problem up so that it conformed to a paradigm he knew, in this case the Walrasian paradigm of his own doctoral dissertation. A more rigorous proof of the existence of the equilibria Fisher was looking for came much later, from Arrow and Debreu (1954). As we know, the problems of infinity, whether agents are assumed to have infinite or finite horizons, are much more troublesome than Fisher imagined.

In any event, Fisher had an excellent vantage point from which to comment on the controversies over capital and interest raging in his day. His formulation of 'investment opportunities' seems to allow for no factor of production one could call 'capital' and enter as argument in a production function. For that matter, he doesn't explicitly model the role of labour in production either, or of land. Strangely, in Fisher's insistence that interest is *not* a cost of production, he seems to say that labour is the only cost, evidently because labour and labour alone is a source of disutility, the loss of utility from leisure, the opportunity cost of the consumption afforded by work. Proceeding in the same spirit, he postulates that, from a position of equality of present and planned future consumption a typical individual will require more extra future consumption than present consumption as compensation for extra work. The difference, the agio, is interest, whether or not it is a 'cost'. Fisher attributes the agio to 'impatience', at the same time scorning the notion that interest is the cost of securing the services of a factor of production called 'abstinence' or 'waiting'.

In the 1890s and 1900s Knut Wicksell, discovering marginal productivity independently of Clark, was modelling production as a function of labour and land inputs with the output also depending on the lags between those inputs and the harvests (Wicksell [1911], 1934, vol. I, pp. 144–66). This is an 'Austrian' formulation, akin to Böhm-Bawerk's examples of trees and wine, in which time itself appears to be productive. Fisher rightly objects to any generalization that waiting longer increases output. His own intertemporal frontiers are, to be sure, sufficiently general to encompass such technologies. They can also accommodate Leontief inputoutput tables and Koopmans-Dantzig activity matrices with lags, Hayekian triangular structures with inventories of intermediate goods in process, Solow technologies with durable goods and labour jointly yielding output contemporaneously or later. The only common denominator of these and other representations of technology is that they relate consumption opportunities at different dates to one another, though not necessarily always in the convex tradeoff terms Fisher assumed. There does not appear to be any summary scalar measure to which the productivity of a process is generally monotonically related, whether roundaboutness, average period of production, or replacement value of existing stocks of goods.

Fisher describes himself as an advocate of 'impatience' as an explanation of interest, although he realizes there are two sides of the saving-investment market, and although he acknowledges that real interest rates can at times be zero or negative. He does appear to believe that in a stationary equilibrium with constant consumption streams, consumers will require positive interest, and that only those technologies and investment opportunities affording a 'rate of return over cost' equal to this pure time preference rate would be used. He does not face up to Schumpeter's argument in 1912 that in such a repetitive and riskless 'circular flow', rational consumers would not care whether a marginal unit of consumption occurs today or tomorrow (Schumpeter [1912], 1934, pp. 34–6). Like Böhm-Bawerk, Fisher appeals to the shortness and uncertainty of life as a reason for time preference. For life-cycle consumers, however, time preferences are entangled with age preferences, and it is hard to defend any generalization as to their net direction. Fair annuities take care of the uncertainty.

Monetary Theory: The Equation of Exchange and the Quantity Theory

Irving Fisher was the major American monetary economist of the early decades of this century; the subject occupied him until the end of his career. Here especially Fisher combined theorizing with empirical research, both historical and statistical. The problems he encountered led him to invent statistical and econometric methods – index numbers and distributed lags in particular – to apply for the purposes at hand to the data he and his assistants compiled. (He even studied the turnover of cash and checking accounts of a sample of Yale students, professors and employees.)

Money was a big subject in American economic literature in the 19th century, before Fisher came on the scene. The monetary events of the times – the inconvertible greenbacks issued during the Civil War, their redemption in gold in 1879, the demonetization of silver, the rapidly increasing importance of banks – stimulated research and controversy. Nevertheless, monetary theory was relatively undeveloped and unsystematized, both in Europe and in America. Fisher's treatise (1911a) was an ambitious attempt to organize with the help of theory a large body of historical and institutional information.

Yet for all its theory, statistics and index numbers, *The Purchasing Power of Money* is a tract supporting Fisher's proposal for stabilizing the value of money. This came to be known as the 'compensated dollar', the gold-exchange standard combined with a rule mandating periodic changes in the official buying and selling prices of gold inverse to changes in a designated commodity price index. In 1911 Fisher proposed that the gold price changes be uniform and synchronous in the currencies of all countries linked by fixed exchange parities, in proportional amounts related to an international price index. Later he was willing to accept as second best that the United States adopt the scheme on its own. Keynes proposed a similar but less formal rule for the United Kingdom (1923).

The proposal is an early example of a policy *rule*, another Fisherian idea ahead of its time, more likely to be popular among economists today than it was with Fisher's contemporaries. Indeed, some rules recently proposed are quite Fisherian, for example Hall (1985).

The 'compensated dollar' is but one of several proposals Fisher advanced over the years for stabilizing price levels or mitigating the effects of their unforeseen variation. In the 1911 book he also writes favourably of the 'tabular standard', which meant no more operationally than facilitating priceindexed indexed contracts. In the 1920s he launched a crusade for 100 per cent reserves against checkable deposits, culminating in *100% Money* (1935). This idea is also beginning to resurface in the 1980s as a preventive defence against the monetary hazards of bank failures. In Schumpeter's view, Fisher's zeal for monetary reforms lost him some of the attention and respect his scientific contributions to monetary economics deserved, and made him come across as more monetarist than his own analysis and evidence justified (Schumpeter 1954, pp. 872–3).

The Purchasing Power of Money is a monetarist book. Fisher asserts the quantity theory as earnestly and persuasively as Milton Friedman. There are two species of quantity theories. One is a simple implication of the 'classical dichotomy': since only relative prices and real endowments enter commodity and factor demand and supply functions, the solution values for real variables in a general equilibrium are independent of scalar variations of exogenous nominal quantities. While Fisher mentions this implication of general equilibrium theory, he does not dwell upon it as one might expect. Anyway, it does not quite apply to a commodity money system like the gold standard, which Fisher was analysing. Fisher's theory is mainly of the second kind, based on the demand for and supply of the particular nominal assets serving as media of exchange.

Fisher is usually given credit for the Equation of Exchange, although Simon Newcomb, a celebrated figure in American astronomy as well as an economist, had anticipated him (1886, pp. 315–47). The Equation is the identity $MV = PT$, where M is the stock of money; V its velocity, the average number of times per year a dollar of the stock changes hands; P is the average price of the considerations traded for money in

such transactions; and T is the physical volume per year of those considerations. It is an identity because it is in principle true by definition. Actually Fisher, of course, recognized the heterogeneity of transactions by writing also $MV = \sum p_i Q_i$, where the p_i and Q_i are individual prices and quantities. His interest in index numbers was substantially a quest for aggregate indexes P and T derived from the individual p_i and Q_i in such a way that the two forms of the equation would be consistent. Much of the book (1911a), both text and technical appendices, is devoted to this quest.

Here and in later writings, particularly (1921) and (1922), Fisher was looking for the 'best' index number formula. He postulated certain criteria and evaluated a host of formulas, investigating their properties both *a priori* and from applications to data. Since the criteria inevitably conflict, there can be no formula that excels on all counts. Although Fisher was mainly interested in measuring movements of the aggregate price level, naturally he wanted a price index P and a quantity index T to have the property that $P_1T_1/P_0T_0 = (\sum p_1Q_1)/(\sum p_0Q_0)$, where the subscripts represent two time periods at which observations of p's and Q's are available.

This and various other desirable consistency properties are not hard to meet. The difficult question is the choice of weights in the two indexes, especially when a whole series of consistent period-to-period comparisons is desired, not just one isolated comparison. For a price index, should the quantity weights be those of a fixed base year, yielding what we now call a 'Laspeyres' index $(\sum p_1Q_0)/(\sum p_0Q_0)$? Or should the weights be those of the ever-changing current period, yielding a 'Paasche' index $(\sum p_1Q_1)/(\sum p_0Q_1)$? The indicated correlate quantity indexes would be the opposites, respectively 'Paasche' and 'Laspeyres'. In 1911 Fisher opted for the Paasche price index. He also seemed to approve the idea of chain indexes, in which the period 0 of the above formulas is not fixed in calendar time but is always the prior period, even though these violate one possible desideratum, that the relative change between two periods should be independent of the base used. He also wrote favourably of the practical advantages of an entirely different procedure, namely taking the median of an expenditure-weighed distribution of percentage price changes from one period to the next.

In 1920, however, Fisher proposed as the 'Ideal Index' a candidate he had not ranked high in 1911, namely the geometric mean of the Laspeyres and Paasche formulas. This formula has the pleasant property that the correlate of an Ideal price index is an Ideal quantity index. Correa Walsh, another index number expert, on whose comprehensive treatise (1901) Fisher relied heavily from the beginning of his own investigations, reached the same conclusion independently at about the same time (Walsh 1921).

These index number issues do not seem as important to present-day economists as they did to Fisher. Knowing that they are intrinsically insoluble, we finesse them and use uncritically the indexes that government statisticians provide. But Fisher's explorations have been important to those practitioners.

In Fisher's Equation of Exchange (1911a) the T and the Q_i are measures of all transactions involving the tender of money, intermediate goods and services as well as final goods and services, old goods as well as newly produced commodities, financial assets as well as goods. The corresponding velocity is likewise comprehensive, much more so than the 'income' or 'circuit' velocity preferred by some monetary theorists, notably Alfred Marshall and his followers in Cambridge (England), who count only transactions for final goods, for example for Gross National Product.

Fisher elaborated the equation to distinguish the quantities M and M' of the two media currency and checking deposits and their separate velocities V and V' $MV + M'V' = PT$. This was a bow to the rising importance of bank deposits relative to currency as transactions media. Previous practice counted only government-issued currency as money, in modern parlance high-powered or base money, and regarded bank operations as increasing its velocity rather than adding to a money stock.

How does the quantity theory come out of the Equation of Exchange? Fisher argues that the real volume of money-using transactions T is

exogenous; that the velocities are determined by institutions and habits and are independent of the other variables in the equation; that the division of the currency supply, the monetary base in current terminology, between currency and bank reserves is stable and independent of the variables in the equation; that banks are fully 'loaned up' so that deposits M' are a stable multiple of reserves, determined by the prudence of banks and by regulation; that exogenous changes in currency supply itself are the principal source of shocks, which, given the preceding propositions, move price level P proportionately. The many qualifications for transitional adjustments are conscientiously presented, but the monetarist message is loud and clear.

The argument is familiar to modern readers, but certain features deserve notice:

(1) Fisher gives the most illuminating account available of the institutions and habits that generate the society's demand for transactions media relative to the volume of transactions. He rightly emphasizes the fact that, and the degree to which, receipts and payments are imperfectly synchronized. He seeks the determinants of velocity in such features of social and economic structure as the frequency of wage and bill payments and the degree of vertical integration of firms. His belief that these institutions change only slowly supports his contention that velocities are exogenous constants.

(2) Much ink has been spilled on the difference between Fisher's velocity approach to money demand and the Cambridge (England) 'k' formulation. The latter, like Walras's *encaisse desiré,* directs attention to agents' portfolio decisions. To Fisher's critics that seems behavioural, while velocity is mechanical. The issue is overblown; the same phenomena can be described in either language. If the other variables in the equation are defined and measured the same way, then V and k are just reciprocals each of the other. Fisher himself discusses hoarding. Fisher's explicit attention, in discussing economy-wide demand for circulating media in distinction to other stores of value, to the fact that money 'at rest' soon takes 'wing' to fly from one agent to another seems to be a merit of his approach.

(3) As already noted, Fisher resolved a question current in his day, whether banks' creation of deposit substitutes for currency should be regarded as increasing the velocity of basic money or as enlarging the supply of money. His choice of the latter course compels attention to the structure, behaviour and regulation of banks. He could not be expected to foresee that the proliferation of future candidates for designation as 'money' would create the monetarist ambiguities we see today.

(4) For the most part later writers have not followed Fisher in his preference for a comprehensive concept and measure of transactions volume. It is hard to attach meaning to the *real* volume of financial transactions, and therefore to see why a T that includes them should be a constant or exogenous term in the equation. On the other hand, modern students of money demand tend simply to forget transactions other than those on final payments.

(5) Fisher ignores the possibility that other liquid assets can serve as imperfect substitutes for money holdings because they can be converted into means of payment as needed, though at some cost. Partly for this reason, he ignores interest rate effects on demand for transactions media. In his day there may have been more excuse for these omissions than there was later. But they are still surprising for an author who elsewhere pays so much attention to the effects of interest rates and opportunity costs on behaviour.

(6) When Fisher was writing, the United States was on the gold standard; the exchange parities of the dollar with sterling and other gold-standard currencies were fixed. Fisher discusses in detail the implications of foreign transactions for the elements of the Equation of Exchange and for the quantity theory. He recognizes that tendencies towards purchasing-power parity, even though imperfect, make money supplies in any one country endogenous, tie prices to those of other

countries and enhance quantity adjustments to monetary shocks in the short run. Much of the 1911 book applies, therefore, to the gold standard economies in aggregate. Indeed, Fisher finds the increase in gold production after 1896 to be the main cause of price increases throughout the world.

Macroeconomics: Business Fluctuations and the Great Depression

The quantity theory by no means exhausts Fisher's ideas on macroeconomics. His views were much more subtle then straightforward monetarism, but they are scattered through his writings and not systematically integrated. Consider the following non-neutralities emphasized by Fisher:

(1) Probably Fisher's principal source of fame, especially among non-economists, is his equation connecting nominal interest i, real interest r and inflation π: $i = r + \pi$. It is frequently misused. Like the Equation of Exchange, it is first of all an identity, from which, for example, an unobservable value of r can be calculated from observations of the other two variables. More interesting, certainly to Fisher, is its use as a condition of equilibrium in financial markets; for this purpose π must be replaced by expected inflation π^e, another unobservable. In a longer run, as Fisher recognized, steady-state equilibrium would also be characterized by equality of actual and expected inflation: $\pi = \pi^e$.

The Fisher equation is frequently cited nowadays in support of complete and prompt pass-through of inflation into nominal interest rates. Fisher's view throughout his career was quite different. For one thing, neither Fisher's theory of interest nor his reading of historical experience suggested to him that equilibrium real rates of interest should be constant. Moreover, from (1896) on he believed that adjustment of nominal interest rates to inflation takes a very long time. This he confirmed by sophisticated empirical investigations, regressions in which the formation of inflation expectations was modelled by distributed lags on actual inflation. During the transition, inflation would lower real rates; nominal rates would adjust incompletely. The effect was symmetrical; he attributed the severity of the Great Depression to the high real rates resulting from price deflation.

Moreover, Fisher was quite explicit about the effects of these movements of real interest rates on real economic variables, including aggregate production and employment. In *The Purchasing Power of Money* these transitional effects are mentioned, but minimized in the author's zeal to convince readers of the importance of stabilizing money stocks. But in Fisher's writings on interest rates, the transitions turn out to be long. In his accounts of cyclical fluctuations in business activity, and especially of the Great Depression, they play the key role.

(2) An assiduous student of price data, Fisher knew that some prices were more flexible than others, that money wages were on the sticky side of the spectrum, and that the imperfect flexibility of the price level meant that the T on the right-hand side of his Equation of Exchange would absorb some of the variations of the left-hand side.

In the early 1930s he came to a very modern position. Real variables like production and employment are independent of the level of prices, once the economy has adjusted to the level. But they are not independent of the rate of change of prices; they depend positively on the rate of inflation. He even calculated a 'Phillips' correlation between employment and inflation (1926). He was just one derivative short of the accelerationist position (Friedman 1968); in a little more time he would have made that step, aware as he was of the difference between actual and expected inflation. Anyway, his policy conclusion was that stabilizing the price level would also stabilize the real economy.

(3) During the Great Depression, observing the catastrophes of the world around him, which he shared personally, Fisher came to quite a different theory of the business cycle from the

simple monetarist version he had espoused earlier. This was his 'debt-deflation theory of depression' (1932), summarized in the first volume of *Econometrica,* the organ of the international society he helped to found (1933). The essential features are that debt-financed Schumpeterian innovations fuel a boom, followed by a recession which can turn into depression via an unstable interaction between excessive real debt burdens and deflation. Note the contrast to the Pigou real balance effect, according to which price declines are the benign mechanism that restores full-employment equilibrium. The realism is all on Fisher's side. This theory of Fisher's has room for the monetary and credit cycles of which he earlier complained, and for the perversely pro-cyclical real interest rate movements mentioned above.

Fisher did not provide a formal model of his latter-day cycle theory, as he probably would have done at a younger age. The point here is that he came to recognize important non-monetary sources of disturbance. These insights contain the makings of a theory of a determination of economic activity, prices, and interest rates in short and medium runs. Moreover, in his neoclassical writings on capital and interest Fisher had laid the basis for the investment and saving equations central to modern macroeconomic models. Had Fisher pulled these strands together into a coherent theory, he could have been an American Keynes. Indeed the 'neoclassical synthesis' would not have had to wait until after World War II. Fisher would have done it all himself.

His practical message in the early 1930s was 'Reflation!' When his Yale colleagues and orthodox economists throughout the country protested against public-works spending proposals and denounced Roosevelt's gold policies, Fisher was a conspicuous dissenter. He was right. Characteristically, he crusaded vigorously for his cause – in speeches, pamphlets, letters and personal talks with President Roosevelt and other powerful policy-makers. Characteristically too, as his letters home (I.N. Fisher 1956, p. 275) disclose, he saw clearly and unapologetically that in lobbying for what was good for the country he was also hoping to rescue the Fisher family finances.

Addressing the President of Yale shortly after Fisher's death, Joseph Schumpeter and eighteen colleagues in the Harvard economics department wrote, 'No American has contributed more to the advancement of his chosen subject. The name of that great economist and American has a secure place in the history of his subject and of his country.' According to his son, this is the eulogy that would have pleased Irving Fisher the most (I.N. Fisher 1956, pp. 337–8). Today, four decades later, economists can confirm the judgement and prediction of that eulogy.

Author's Note: Fortunately Fisher's son, Irving Norton Fisher, preserved the memory of his father in two indispensable publications, a biography and a comprehensive bibliography (1956, 1961). I have also relied extensively on Professor John Perry Miller's biographical essay (1967) and Professor William Barber's account (1986) of political economy at Yale before 1900. My review of Fisher's contributions to general equilibrium theory, the theory of capital and interest, monetary theory and macroeconomics draws heavily and often literally on a recent essay of my own (Tobin 1985).

Selected Works

1892. *Mathematical investigations in the theory of value and prices.* New Haven: Connecticut Academy of Arts and Sciences, *Transactions 9,* 1892. Reprinted, New York: Augustus M. Kelley, 1961.

1896. Appreciation and interest. *AEA Publications* 3(11): 331–442. Reprinted, New York: Augustus M. Kelley, 1961.

1906a. *The nature of capital and income.* New York: Macmillan.

1906b. *The rate of interest.* New York: Macmillan.

1910. *Introduction to economic science*. New York: Macmillan.

1911a. *The purchasing power of money.* New York: Macmillan.

1911b. *Elementary principles of economics*. New York: Macmillan.
1921. The best form of index number. *American Statistical Association Quarterly* 17: 533–537.
1922. *The making of index numbers*. Boston: Houghton Mifflin.
1926. A statistical relation between unemployment and price changes. *International Labour Review* 13: 785–792.
1927. A statistical method for measuring 'marginal utility' and testing the justice of a progressive income tax. In *Economic essays contributed in honor of John Bates Clark*, ed. J.H. Hollander. New York: Macmillan.
1930. *The theory of interest*. New York: Macmillan.
1932. *Booms and depressions*. New York: Adelphi.
1933. The debt-deflation theory of great depressions. *Econometrica* 1(4): 337–357.
1935. *100% money.* New York: Adelphi.

Bibliography

Arrow, K.J., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22(3): 265–290.
Auspitz, R., and R. Lieben. 1889. *Untersuchungen über die Theorie des Preises*. Leipzing: Duncker & Humblot.
Barber, W.J. 1986. Yale the fortunes of political economy an environment of academic conservatism. In *Economists and American higher learning in the nineteenth century*, ed. W.J. Barber. Middletown: Wesleyan University Press.
Clark, J.B. 1899. *The distribution of wealth*. New York: Macmillan.
Fairchild, F.R., Furniss, E.S., and Buck, N.S. 1926. *Elementary economics*, 2 vols. New York: Macmillan. 5th ed, 1948.
Fisher, I.N. 1956. *My father Irving Fisher*. New York: Comet Press.
Fisher, I.N. 1961. *A bibliography of the writings of Irving Fisher*. New Haven: Yale University Library.
Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58(1): 1–17.
Hall, R.E. 1985. Monetary policy with an elastic price standard. In *Price stability and public policy*, 137–160. Kansas City: Federal Reserve Bank of Kansas City.
Jevons, W.S. 1871. *The theory of political economy.* London: Macmillan.
Keynes, J.M. 1923. *A tract on monetary reform*. London: Macmillan.
Keynes, J.M. 1936. *The general theory of employment, interest and money.* New York: Harcourt, Brace.
Miller, J.P. 1967. Irving Fisher of Yale. In *Ten economic studies in the tradition of Irving Fisher*, ed. William Fellner et al. New York: Wiley.
Newcomb, S. 1886. *Principles of political economy.* New York: Harper.
Rae, J. 1834. *The sociological theory of capital*. Reprinted. New York: Macmillan, 1905.
Samuelson, P.A. 1967. Irving Fisher and the theory of capital. In *Ten economic studies in the tradition of Irving Fisher*, ed. William Fellner et al. New York: Wiley.
Scarf, H. (With T. Hansen.) 1973. *The computation of economic equilibria.* New Haven: Yale University Press.
Schumpeter, J.A. 1912. *Theory of economic development.* Trans. from the 2nd German edn of 1926 by R. Opie. Cambridge, MA: Harvard University Press, 1934.
Schumpeter, J.H. 1954. *History of economic analysis.* Ed. E.B. Schumpeter. New York: Oxford University Press.
Social Sciences Citation Index. 1979, 1983. *Five year cumulation*, 1971–5 and 1976–80. Philadelphia: Institute for Scientific Information.
Tobin, J. 1985. Neoclassical theory in America. *American Economic Review* 75(6): 28–38.
Walsh, C.M. 1901. *The measurement of general exchange value*. New York/London: Macmillan.
Walsh, C.M. 1921. *The problem of estimation*. London: King & Sons.
Wicksell, K.M. 1911. *Lectures on political economy.* Trans E. Classen from the 2nd Swedish edn. London: George Routledge & Sons, 1934.

Fisher, Ronald Aylmer (1890–1962)

A.W.F. Edwards

Keywords

Ancillarity; Conditional inference; Fiducial probability; Fisher, R. A.; Game theory; Likelihood; Natural selection; Statistical inference

JEL Classifications

B31

R.A. Fisher was born in London on 17 February 1890, the son of a fine-art auctioneer. His twin brother was stillborn. At Harrow School he distinguished himself in mathematics, despite being handicapped by poor eyesight which prevented him working by artificial light. His teachers used to instruct by ear, and Fisher developed a remarkable capacity for pursuing complex mathematical arguments in his head. This manifested itself in later life in his ability to reach a conclusion whilst forgetting the argument; to handle complex geometrical trains of thought; and to develop and report essentially mathematical arguments in English (only for students to have to reconstruct the mathematics later).

He entered Gonville and Caius College, Cambridge, as a scholar in 1909, graduating BA in mathematics in 1912. Prevented from entering war service in 1914 by his poor eyesight, Fisher held several jobs before being appointed Statistician to Rothamsted Experimental Station in 1919. In 1933 he became Galton Professor of Eugenics at University College London, and in 1943 Arthur Balfour Professor of Genetics in Cambridge and a Fellow of Caius College. He retired in 1957 and spent his last few years in Adelaide, Australia, where he died from a post-operative embolism on 29 July 1962.

He married Ruth Eileen Guiness in 1917 and they had two sons and six daughters. He was elected a Fellow of the Royal Society in 1929 and was knighted in 1952 for services to science.

Fisher made a most profound contribution to applied and theoretical statistics and to genetics. He had been attracted to natural history, and especially the works of Darwin, at school, and he had bought Bateson's *Principles of Genetics,* with its translation of Mendel's paper, in his first term as an undergraduate. Before graduating he had already remarked on the surprisingly good fit of Mendel's data, published a paper introducing the method of maximum likelihood, and given a proof of the distribution of the *'t'* statistic which Student had only conjectured.

In 1915 Fisher published the distribution of the correlation coefficient; in 1918 the seminal work in biometrical genetics, 'The correlation between relatives on the supposition of Mendelian inheritance', in which he introduced the word 'variance' and foreshadowed his later development of the analysis of variance; and in 1922 'On the Mathematical Foundations of Theoretical Statistics', a paper which revolutionized statistical thought.

As Statistician at Rothamsted he founded the subject of experimental design based on randomization, pursued vigorously the development of statistical estimation theory and invented – or, at least, captured – the quixotic notion of fiducial probability. Moving to London the pace did not slacken, for in addition to pioneering genetical work, especially in connection with the human blood groups, Fisher's statistical explorations revealed the likelihood principle, conditional inference and the concept of ancillarity.

The Second World War found him embattled on many fronts. Unhappy at home, he found his scientific activity disrupted by wartime conditions including the evacuation of his department from London. The profundity of his work on statistical inference was ill-appreciated in America, where preoccupation with wartime problems encouraged an excessively mathematical and operational view with which Fisher had little sympathy. In mathematical genetics there were similar difficulties as the American school, starting from his 'fundamental theorem of natural selection', developed ideas of 'adaptive topographies' with false analogies to physical systems. It was not until well after his death that in both statistical inference and mathematical genetics the criticisms which he had advanced came to be appreciated.

After the war, from the relative peace of Cambridge, Fisher saw his theoretical work in both subjects suffer further temporary eclipse. He made great, but ultimately unsuccessful, efforts to establish biochemical genetics in his department and to secure for Cambridge the national laboratories for human blood-group work. When close to retirement, he was amongst the first to realize the significance of Watson and Crick's discovery of the structure of DNA (1953), and to apply the new computers to a biological problem (1950).

Perhaps embittered by his post-war experiences (though he never relaxed his scientific

work), he found some consolation in the Presidency of Caius College from 1956 to 1959, a post second to the Master, and further happiness in retirement in Adelaide.

Fisher wrote five books and published a famous set of statistical tables jointly with F. Yates. An extremely informative and admirably objective biography was published by one of his daughters in 1978 (Box 1978).

In the field of economics Fisher's name would be remembered for his contributions to statistics alone, so fully chronicled in Box's biography, but we may here draw attention to three other areas not emphasized in the biography but which are especially relevant.

First, the 'fundamental theorem of natural selection' (1930). Although this is specifically directed at a genetical problem, it relies on a simpler implicit theorem of widespread relevance wherever discussion centres on differential growth rates, namely 'the rate of change in the growth-rate is proportional to the variance in growth-rates'. This precise theorem, which is easily proved mathematically, captures the notion that the growth rate of the fastest-growing sub-population (or economic sector, and so on) will come to dominate the overall growth rate.

Secondly, the modern preoccupation with 'socio-biology' has as one of its origins *The Genetical Theory of Natural Selection* (1930), a fact that only surprises those who have not studied the book and Fisher's other writings on human affairs in the two decades before the Second World War.

Thirdly, Fisher not only introduced the theory of games into evolutionary biology (at the suggestion of Dr Cavalli, later Professor Cavalli-Sforza), but he discovered and published the idea of a randomized or 'mixed' strategy as early as 1934, independently of von Neumann. The problem was the card game *'Le Her'*, though if Fisher had gone to the primary source (the correspondence between Montmort and Nicholas Bernoulli, published in 1713) rather than relying only on Todhunter's *History of the Mathematical Theory of Probability* (1865), he would have found that his solution had already been given by Waldegrave.

Selected Works

1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10: 507–521.

1918. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52: 399–433.

1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions. Royal Society of London, Series A* 222: 309–368.

1925. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

1930. *The genetical theory of natural selection*. Oxford: Clarendon Press.

1935. *The design of experiments*. Edinburgh: Oliver and Boyd.

1938. With F. Yates. *Statistical tables for biological, agricultural and medical research.* Edinburgh: Oliver and Boyd.

1949. *The theory of inbreeding.* Edinburgh: Oliver and Boyd.

1950. *Contributions to mathematical statistics*. New York: Wiley.

1956. *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.

1971–4. *Collected papers of R.A. Fisher*, 5 vols, ed. J.H. Bennett. Adelaide: University of Adelaide.

Bibliography

Box, J.F. 1978. *R.A. Fisher: The life of a scientist*. New York: Wiley.

Bibliographic Addendum

A celebrated discussion of Fisherian statistics is L.J. Savage, 'On Rereading R.A. Fisher (with discussion)', *Annals of Statistics* 4 (1976), 441–500. Other interesting writings include B. Efron, 'R.A. Fisher in the 21st Century', *Statistical Science* 31(2) (1998), 95–122, which demonstrates the continuing direct relevance of Fisher's work; C.R. Rao, 'R.A. Fisher: The Founder of Modern Statistics', *Statistical Science* 7 (1992), 34–48, an appreciation by a former student; and S. Stigler, 'Fisher in 1921', *Statistical Science* 20 (2005), 32–49, which discusses the origins of Fisher's monumental 1922 paper.

Fisheries

Colin W. Clark

Abstract

Marine fisheries throughout the world continued to be severely overexploited throughout the 20th century and beyond. Even under intensive 'scientific' management many important fisheries have collapsed, some never to recover. Vast overcapacity of fishing fleets is also widespread. Both outcomes can be attributed to the common-pool aspect of fishery resources. One method of countering these developments, individual transferable catch quotas (ITQs), is currently in use in several countries. Provided this instrument is combined with resource taxes (royalties), an efficient and equitable management system is feasible. (Owing to lack of jurisdiction, deep-sea fisheries seem destined to continue to suffer from overfishing).

Keywords

Bionomic equilibrium; Common property resources; Extended fisheries jurisdiction (EFJ) zones; Fisheries; Individual fishing quotas (IFQs); Individual transferable quotas (ITQs); Maximum sustainable yield (MSY) paradigm; Precautionary management; Resource rents; Royalities

JEL Classifications

E22

By the end of the 1970s, most of the world's coastal states had declared 200- nautical-mile zones of extended fisheries jurisdiction (EFJ zones) over marine fisheries. These zones allowed coastal states to exert full control over fishing activities. Over 90 per cent of global marine fishery landings thus came under the control of coastal states.

The need to regulate fishing activities arises from the common-pool nature of marine fish (and other living-resource) populations. In simple terms, under unregulated open-access conditions any fish stock that can be profitably harvested will in fact be exploited. Whether such exploitation eventually leads to biological depletion and reduced harvest levels depends on a number of circumstances, including demand for the product, cost of fishing, and the distribution, abundance and behaviour of the fish.

Using a simple graphical model, H.S. Gordon (1954) argued that an unregulated fishery would achieve 'bionomic' equilibrium, reaching a stock level at which the revenue from catching and selling fish would just balance the opportunity costs of fishing. Populations with high price–cost ratio would thus be heavily exploited, while those with low ratio would remain lightly exploited or unexploited. Countless actual examples support this prediction.

The next obvious question is whether bionomic equilibrium is undesirable and, if so, what can be done about it. Early commentators largely agreed that, because bionomic equilibrium results in the dissipation of economic rents, it is economically undesirable, independent of any biological consequences. But in many cases bionomic equilibrium also implies biological overfishing, defined as the reduction of the fish population to a level at which net annual biological productivity is below the maximum that could be generated. Indeed, in extreme cases overfishing can lead to the collapse of the fishery, with little or no recovery after fishing is terminated (Dulvy et al. 2003). Pauly et al. (1998) and Myers and Worm (2003) and other scientists have documented the historical decline of marine fish stocks on a worldwide scale, especially over recent decades. Even many stocks within 200-mile EFJ zones have continued to be overfished. The reasons for this outcome are only beginning to be generally understood.

Fisheries management has traditionally been based on the objective of determining and achieving the maximum sustainable yield (MSY) available from each population. In its own right, this approach is beset with difficulties generated by unobservability and uncertainty pertaining to marine populations and ecosystems (Caddy and Seijo 2005). Management difficulties also arise

because the MSY paradigm overlooks all economic aspects of fishing.

To be more specific, until recently most fishery management programmes have been based almost exclusively on the total allowable catch (TAC) method. The annual TAC is calculated on the basis of an accepted model, and the fishery is managed so as to achieve this quota, usually by means of restricted annual openings of the fishery. If correctly calculated and implemented, this method can indeed prevent overfishing and produce positive economic rents – temporarily.

But *positive rents attract additional fishing effort* – this is the basis of Gordon's original theory. If annual effort is controlled through TAC-based seasonal openings, the response is that either the fishermen increase the power and capacity of their vessels, or additional vessels enter the fishery, or both. Further shortenings of the fishing season are then needed, and so on. A new regulated bionomic equilibrium is reached when the average capital costs of expanding fishing capacity are just equal to the average present value of net operating revenues. Rents in a TAC-managed fishery are then dissipated through over-expansion of fishing capacity rather than via overfishing. Extreme overcapacity of fishing fleets worldwide is today considered to be a major impediment to rational management.

It seems natural, therefore, to attempt to control fishing capacity. In cases where excess capacity has already developed, vessel buy-back programmes have often been used to reduce fleet size. However, such buy-back programmes, which can be very costly, do nothing to eliminate the incentives for additional expansion. Indeed, if buy-backs are anticipated by fishermen, they may actually induce a higher level of initial overcapacity than would otherwise occur (Clark et al. 2005).

Two possible approaches to resolving the joint problems of overfishing and overcapacity are, first, taxes or royalties, and second, individual fishing quotas (IFQs). Although these are usually considered as alternatives, they can in fact readily be used in combination. By the early 21st century IFQs and related programmes were in effect in several countries, with generally positive results (Cunningham and Bostock 2005; Clark 2006).

IFQs can be envisioned as a form of quasi-property rights, an interpretation that is strengthened if the quotas are tradable (that is, individual transferable quotas, ITQs). Its owner considers an ITQ as a productive asset, whose value will be enhanced if the resource is protected and well-managed. It also seems likely (though this remains to be demonstrated in practice) that ITQ owners will favour risk-averse management strategies, such as conservative TACs and the use of marine reserves.

Various economic distortions can arise, however, if ITQs are awarded free of charge. For example, the initial recipients of the quotas may become greatly enriched. This possibility being well known to fishermen, the anticipation of a forthcoming ITQ programme may attract extra entry into the fishery, dissipating much of the future rents in advance. Besides this there is the question of social equity – why should the government award special access privileges to a publicly owned resource to a chosen few individuals? Charging significant catch royalties can reduce rent-seeking incentives, while also compensating the resource owner, namely, the general public.

During the current transitional phase from managing ocean fisheries as common-pool resources to managing them with individual quotas, royalty charges will probably remain minimal. But, once a profitable fishery develops, it seems likely that the public will expect and demand a fair share of the resource rents – as is already the case with other natural resource assets.

Whatever system is used, the management of marine fisheries will always face high levels of uncertainty. Marine ecosystems are complex, poorly observable, and subject to unpredictable, environmentally induced fluctuations. Finely tuned management, intended for example to maximize some specified objective, will remain elusive. The threat of overfishing persists, even for closely monitored and managed stocks. Also, recent experience has shown that the recovery of depleted stocks can often be slow or non-existent (Hutchings 2000).

For these reasons it is now widely agreed that a precautionary management approach is needed (Charles 2001). Conservative annual catch quotas are necessary to protect against inadvertent

overfishing. In addition, breeding stocks need to be strongly protected, as do sea-floor and estuarine habitats, and marine ecosystems in general. Furthermore, fishing activities that damage and degrade the marine environment, leading to long-term reductions in productivity, need to be controlled or eliminated.

Marine reserves, permanently protecting substantial areas of the ocean from harvesting activities, can provide a valuable hedge against management error resulting from biological uncertainty or from imperfect control of fishing operations. Such reserves can protect breeding stocks, ensuring a continued supply of recruits even when stocks are overfished elsewhere. Reserves are not a substitute for well-designed and operated traditional management systems; rather, they need to be used in conjunction with normal management methods.

Space limitations preclude the discussion of other important issues such as: ocean pollution, aquaculture, illegal fishing and non-regulated deep-sea fisheries, and ecosystem-based management programmes.

See Also

► Common Property Resources

Bibliography

Caddy, J., and J. Seijo. 2005. This is more difficult than we thought: The responsibility of scientists, managers and stakeholders to mitigate the unsustainability of marine fisheries. *Philosophical Transactions of the Royal Society B* 360: 59–75.

Charles, A. 2001. *Sustainable fishery systems*. Oxford: Blackwell Science.

Clark, C. 2006. *Worldwide crisis in fisheries*. Cambridge: Cambridge University Press.

Clark, C., G. Munro, and U. Sumaila. 2005. Subsidies, buybacks and sustainable fisheries. *Journal of Environmental Economics and Management* 50: 47–58.

Cunningham, S., and T. Bostock, eds. 2005. *Successful fisheries management*. Delft: Eburon Academic Publishers.

Dulvy, N., Y. Sadovy, and J. Reynolds. 2003. Extinction vulnerability in marine populations. *Fish and Fisheries* 4: 24–64.

Gordon, H. 1954. The economic theory of a common property resource: The fishery. *Journal of Political Economy* 62: 124–142.

Hutchings, J. 2000. Collapse and recovery of marine fishes. *Nature* 406: 882–885.

Myers, R., and B. Worm. 2003. Rapid worldwide depletion of predatory fish communities. *Nature* 423: 280–283.

Pauly, D., V. Christensen, J. Dalsgaard, R. Froese, and F. Torres. 1998. Fishing down marine food webs. *Science* 279: 860–863.

Fixed Capital

Paolo Varri

Fixed capital is the term traditionally used to indicate durable means of production, that is all those inputs of the productive process (such as tools, machines and equipment) that are not exhausted is one single period of production. Non-durable means of production, by contrast defined circulating capital, include raw materials, energy, direct labour, semi-finished goods, etc.

Of course, while circulating capital contributes entirely to the annual production of each commodity, the contribution of fixed capital to production in each period should be determined in relation to the wear and tear actually incurred during its utilization; a datum that in general is not possible to observe directly.

Fixed capital is therefore a complication in the theory of production and it is easy to understand the reason why economists, in their search for abstract simplification of very complex real phenomena, are often induced to assume that production requires only circulating capital.

But technical progress has continuously increased the relevance of machines and plant in industrial production and, as a consequence, a theory of production able to face the problem of fixed capital has become more and more necessary. The most interesting recent contribution in this direction does not belong to mainstream traditional neoclassical theory. It has been made by Sraffa (1960) going back to the classical tradition of determining the value of commodities according to their conditions of production.

Historical Developments

Fixed capital is already present in the propositions of the early economists. The determination of its contribution to the annual product of a nation by the Physiocrats and Adam Smith (1766) is however only a description of the behavioural rules of the business world rather than an attempt to explain them. The first analytical discussion of the problem of fixed capital is associated with Ricardo (1821). He is concerned with two particular aspects of the problem.

First of all he noticed that, when the rate of profits is changed, the presence of fixed capital is one of the factors that may alter the proportionality between the ratio of prices and the ratio of the quantity of labour embodied in the corresponding commodities. This is the famous exception of time to the general rule of the labour theory of value that Ricardo put forward in reply to the criticisms raised by McCulloch.

The second aspect of the problem of fixed capital considered by Ricardo, is concerned with the effects of the dynamic substitution in production of machines for labour. He concludes that workers' fears of technological unemployment may be justified, even if the conclusion does not seem to follow logically from his model, that is based on Say' Law.

Marx (1867–1894) analyses in detail the consequences of the introduction of fixed capital (machines) on the productivity of labour and strongly underlines the enormous reduction in the price of commodities that it implies; but apparently he does not care to determine the contribution of fixed capital to the cost of production in each period. A second deeper implication that Marx draws from the substitution in time of machines for labour is the increase in the organic composition of capital, from which he derives his controversial tendency of the rate of profits to fall.

Recent Contributions

There are two distinct contributions that, in very different ways, are relevant for the modern analysis of fixed capital: von Neumann (von 1937) and Leontief (1941); Leontief et al. (1953). Von Neumann spends only few words in describing the economic meaning of his mathematical model, but he explicitly remarks that capital goods should appear in both the input and in the output matrix of his model, and should be considered as different goods for each different stage of their utilization, i.e. exactly the same method of analysis later adopted by Sraffa that, nevertheless, at the moment, did not receive any particular attention.

The second contribution, Leontief's input–output model, is relevant because it has many analogies with Sraffa's scheme of production and because Leontief explicitly tries to introduce fixed capital in his model. This is therefore a good starting point to appreciate Sraffa's solution of the problem.

Leontief's (1941) input–output model is a scheme of the flows of commodities among the various industries of the economic system initially conceived to take into account only circulating capital. It determines the quantities of the commodities produced and their prices as solutions of the following two systems of equations:

$$Aq + y = q \tag{1}$$

$$pA + v = p \tag{2}$$

where A is the input-output matrix of technical coefficients, q and y are the vectors of total production and of final demand, p is the vector of prices and υ is the vector of value added.

But, as Leontief et al. (1953) himself later recognized, a more complete description of the economic system must also involve stocks of commodities (fixed capital) in their various forms: inventories, machines, buildings, etc. He introduces therefore a second square matrix $B = b_{ij}$ that indicates the amount of commodity i required as stock to produce one unit of commodity j. Bq is then the vector of stocks of commodities required to produce the vector of commodities q. Fixed capital stocks affect the balance equation of each period only in terms of the variations of the levels of production $\dot{q} = dq/dt$. This leads Leontief to analyse the dynamic implications of the introduction of fixed

capital by means of the following system of linear differential equations:

$$y = q - Aq - B\dot{q} \tag{3}$$

showing the interaction of stocks and flows as a generalization of the acceleration principle.

Whatever the interest of these dynamic extensions may be, the treatment of fixed capital is rather crude because the determination of depreciations (the fundamental problem with fixed capital) remains exogenous to the model. The amount of fixed capital consumed in each year is in fact predetermined by simplifying assumptions either as a share of the initial stock or as a fixed percentage rate of decay of the residual stock and it is included in the flow matrix A.

Fixed Capital in a General Scheme of Flows

Sraffa's (1960) approach allows a substantial analytical improvement on the problem of fixed capital. He does not consider machines as stocks à *la* Leontief and proposes instead to consider what remains of a machine at the end of each year of operation as a joint product together with the commodity produced. An approach that Sraffa first attributes to Torrens and that afterwards was adopted by Ricardo, Malthus and Marx and then fell into oblivion with the already mentioned exception of von Neumann.

The main interest of Sraffa is in the theory of value and distribution of income. Following the approach of the classical economists, that tried to determine prices from the conditions of production of each commodity, Sraffa formulates a scheme of the production system articulated in two stages. At the first stage of the analysis, when each industry is supposed to produce one single commodity, and the number of industries is equal to the number of commodities produced, Sraffa defines a system of equations that is usually written as follows:

$$a_n w + pA(1 + r) = p. \tag{4}$$

It shows that the structure of the production system, as described by the matrix of technical coefficients $A = a_{ij}$ and by the vector of labour coefficients a_n, together with one of the two distributive variables (e.g., the uniform rate of profits r), is sufficient to determine the structure of the vector of prices p and the second residual distributive variable (for analytical details see Newman 1962 and Pasinetti 1977).

The meaning of these prices has nothing to do with marginal or neoclassical theory. They represent a more fundamental concept: the exchange rates which ensure the reproduction of the economic system.

The introduction of fixed capital requires the second stage of the analysis, where each industry may produce jointly more than one single commodity. The outcome of this method of dealing with fixed capital is a general scheme of flows that avoids the hybrid interplay between stocks and flows of Leontief's solution.

Obviously a scheme of general joint production is much more complicated than single production. But it is not necessary to go into all the intricacies of joint production to analyse fixed capital. Sraffa considers fixed capital as the leading species of the genus of joint products, and this has suggested an analysis of the intermediate stage where fixed capital is the only element of joint production in a system of single product industries.

At this particular intermediate stage a new system of equations substitutes for the previous one:

$$a_n w + pA(1 + r) = pB \tag{5}$$

where $B = b_{ij}$ is a square matrix of outputs that indicates the quantity of each commodity produced and the quantity of old machines, as their joint products, and p is the price vector of the commodities produced, including the price of all old machines at their various ages.

By contrast with the case of single production it might well happen here that, for feasible levels of the rate of profits, some price comes out to be negative, but it is possible to show that, if fixed capital is the only element of joint production of the scheme, then, only the price of old machines might be negative. This has a precise economic meaning:

it is a signal of productive inefficiency. It may be shown that, by correspondingly reducing the years of utilization of the machine, the (productive) efficiency of the system would increase (i.e. it would allow higher wages at the same rate of profits). This means that it is always possible, after a suitable truncation of the period of utilization of the machine, to eliminate all negative prices and to obtain a strictly positive solution. (Further analytical details may be found in the essays by Baldone 1974; Schefold 1974 and Varri 1974.)

The method of joint production therefore leads to prices that are economically meaningful and at the same time makes it possible to determine the most efficient life time of durable means of production that turns out to depend, not necessarily in a monotonic way, on the rate of profits.

The remarkable consequence of this result is that, by considering the difference of the prices of the same machine at two subsequent years, it is always possible to obtain the *correct* depreciation quota for that machine in the year considered; correct in the sense of allowing the replacement of the means of production and the payment of profits, whatever the technical conditions of use of the machine may be over its period of utilization. A solution therefore to the problem of determining the wear and tear actually occurred during the utilization of the machine that, as was noticed at the beginning, is impossible to observe directly.

Final Remarks

A remarkable property of the analysis of fixed capital outlined so far is that, though avoiding the difficulties of general joint production schemes, it is rather general and comprehensive. It concerns regular systems where machines are used in their natural sequence and it is necessary to assume that at the end of their life their residual value is zero. Moreover trade of old machines among industries producing different commodities is excluded.

But the analysis does take into account two important complementary aspects of the problem of fixed capital. The first concerns the possibility of considering sets of machines jointly utilized in production, as a unique durable means of production, let us call it a plant, avoiding the indeterminacy of the price of each single component.

The second regards the valuation of obsolete machines no longer produced, but still worth using in production, that may be obtained from the computation of quasi-rents according to the same principle that applies to the rent of lands of different qualities.

More complicated schemes of fixed capital utilization are of course possible but should be analysed within the framework of general joint production.

The most important feature of Sraffa's approach to the problem of fixed capital is that, not requiring any change in the fundamental vision of production as a circular process initially adopted to analyse circulating capital, it greatly contributes to establishing it as a general approach for the analysis of modern systems of production that is alternative to marginalism and neoclassical theory.

See Also

- ▶ Capital as a Factor of Production
- ▶ Capital Goods
- ▶ Circulating Capital

References

Baldone, S. 1974. II capitale fisso nello schema teorico di Piero Sraffa. *Studi Economici* 29: 45–106 .Trans. as: 'Fixed capital in Sraffa's theoretical scheme', in Pasinetti (1980), 88–137

Leontief, W. 1941. *The structure of american economy, 1919–1929*. New York: Oxford University Press.

Leontief, W., et al. 1953. *Studies in the structure of american economy*. New York: Oxford University Press.

Marx, K. 1867. *Capital*. Moscow: Progress Publishers 1965–1967.

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

Newman, P. 1962. Production of commodities by means of commodities. *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 98: 58–75.

Pasinetti, L. 1977. *Lectures on the theory of production*. London: Macmillan.

———. 1980. *Essays on the theory of joint production*. London: Macmillan.

Ricardo, D. 1821. In *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.

Schefold, B. 1974. Fixed capital as a joint product and the analysis of accumulation with different forms of technical progress. Mimeo, published in Pasinetti (1980), 138–217.

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

Sraffa, P. (ed). 1951–1973. *The works and correspondence of David Ricardo*. Cambridge: Cambridge University Press.

———. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.

Varri, P. 1974. Prezzi, saggio del profitto e durata del capitale fisso nello schema teorico di Piero Sraffa. *Studi Economici* 29: 5–44 .Trans. as 'Prices, rate of profit and life of machines in Sraffa's fixed capital model', in Pasinetti (1980), 55–87

Fixed Effects and Random Effects

Badi H. Baltagi

Abstract

Unobservable individual effects in panel data models are employed to control for heterogeneity. These can be thought of as random variables that are uncorrelated with the regressors, thus generating a random effects model. Alternatively, these random individual effects are allowed to be completely correlated with the regressors, thus generating a fixed effects model. The choice between these two alternatives is usually settled using a Hausman (Econometrica 46:1251–1271, 1978) test. This article argues that one should interpret a rejection by the Hausman test as a rejection of the random effects model, not necessarily an endorsement of the fixed effects model.

Keywords

Attrition; Autocorrelation; Cross-section data; Fixed effects; Haavelmo, T; Heteroskedasticity; Instrumental variable estimators; Least squares dummy variables (LSDV) model; Multicollinearity; Over-identification; Panel data; Random effects; Sample selection

JEL Classification

C23

One of the major benefits from using panel data as compared to cross-section data on individuals is that it enables us to control for individual heterogeneity. Not controlling for these unobserved individual specific effects leads to bias in the resulting estimates. Consider the panel data regression

$$y_{it} = \alpha + X'_{it}\beta + u_{it} \quad i = 1, \ldots, N; \quad t = 1, \ldots, T \tag{1}$$

with i denoting individuals and t denoting time. The panel data is *balanced* in that none of the observations is missing whether randomly or non-randomly due to attrition or sample selection. α is a scalar, β is $K \times 1$ and X_{it} is the itth observation on K explanatory variables. Most panel data applications utilize a one-way error component model for the disturbances, with

$$u_{it} = \mu_i + \nu_{it} \tag{2}$$

where μ_i denotes the *unobservable* individual specific effect and ν_{it} denotes the remainder disturbance. For example, in an earnings equation in labour economics, y_{it} will measure earnings of the head of the household, whereas X_{it} may contain a set of variables like experience, education, union membership, sex, or race. Note that μ_i is time-invariant and it accounts for any individual specific effect that is not included in the regression. In this case we could think of it as the individual's unobserved ability. The remainder disturbance ν_{it} varies with individuals and time and can be thought of as the usual disturbance in the regression. If the μ_i's are assumed to be *fixed parameters* to be estimated, we get the *fixed effects (FE) model*. If the μ_i's are assumed random variables independent of X_{it} and ν_{it}, for all i and t, we get the *random effects (RE) model*.

For the fixed effects model, the regression equation in (1) becomes

$$y_{it} = \alpha + X'_{it}\beta + \mu_i + v_{it} \tag{3}$$

where the μ_i's can be estimated as coefficients of dummy variables, one for each individual. This model is also known as the least squares dummy variables (LSDV) model. Note that only $(\alpha + \mu_i)$ is estimable and that is why it is sometimes denoted by α_i. For large labour or consumer panels, where N is very large, LSDV regressions like (3) may not be feasible. In this case, one is including $(N-1)$ dummy variables in the regression and therefore inverting a huge matrix of dimension $(N + K)$ rather than $(K + 1)$ as in (1). In addition, this FE regression suffers from a large loss of degrees of freedom, since we are estimating $(N-1)$ extra parameters, and too many dummies may aggravate the problem of multicollinearity among the regressors. In particular, this FE estimator cannot estimate the effect of any time-invariant variable like gender, race, religion which may be of prime interest for the researcher especially in attempting to estimate wage differentials among men and women or whites and non-whites, with other factors held constant. In fact, these time-invariant variables are spanned by the individual dummies in (3) and therefore any OLS regression attempting to estimate (3) will fail, signalling perfect multicollinearity.

Averaging (3) over time yields

$$\bar{y}_{i.} = \alpha + \overline{X'}_i\beta + \mu + \bar{v}_{i.} \tag{4}$$

Subtracting (4) from (3) gives

$$y_{it} - \bar{y}_{i.} = \left(X_{it} - X'_{i.}\right)^0\beta + (v_{it} - \bar{v}_{i.}). \tag{5}$$

One can show that the FE estimator of β (denoted by $\beta\ \tilde{\beta}_{FE}$) obtained from the sometimes infeasible LSDV regression in (3) can be alternatively obtained from the simpler regression given in (5). The latter regression is known as the *within*-regression since it is based on the within variation in the data. Regression (4), which is a cross-section regression, is known as the *between*-regression since it is based on the between variation in the data. If (3) is the true model, FE is the best linear unbiased estimator (BLUE) as long as the remainder disturbances (the v_{it}'s) are independent and identically distributed (i.i.d.)$(0, \sigma^2)$. Of course, here we are assuming that the X_{it}'s are independent of the v_{it} for all i and t. The fixed effects model is deemed appropriate when one is focusing on a specific set of N countries, states, counties, regions or firms. Inference in this case is conditional on the particular N firms, countries or states that are observed. Note that, if T is fixed and $N \to \infty$ as typical in short labour panels, then only the FE estimator of β is consistent; the FE estimators of the individual effects (α_i) are not consistent since the number of these parameters increases as N increases. This is the *incidental parameter problem* discussed by Neyman and Scott (1948) and reviewed more recently by Lancaster (2000). Note that, when the true model is fixed effects as in (3), pooled OLS on (1) yields biased and inconsistent estimates of the regression parameters. This is an omission variables bias because OLS deletes the individual dummies when in fact they are relevant. One could test the joint significance of these dummies, that is, H_0; $\mu_1 = \mu_2 = \ldots = \mu_{N-1} = 0$, by performing an F-test. This is a simple Chow test with the restricted residual sums of squares (RRSS) being that of OLS on the pooled model and the unrestricted residual sums of squares (URSS) being that of the LSDV regression in (3) or equivalently the residual sum of squares from the within-regression in (5). In this case

$$F_0 = \frac{(RRSS - URSS)/(N-1)}{(URSS/(NT - N - K)} \overset{H_0}{\sim} F_{N-1,N(T-1)-K}. \tag{6}$$

One computational caution for those using the within-regression computed from (5). The s^2 of this regression as obtained from a typical regression package divides the residual sums of squares by $NT-K$ since the intercept and the dummies are not included. The proper s^2, say s^{*2} from the LSDV regression in (3), would divide the same residual sums of squares by $N(T-1)-K$.

F

Therefore, one has to adjust the variances obtained from the within-regression by multiplying the variance-covariance matrix by (s^{*2}/s^2) or simply by multiplying by $[NT-K]/[N\ (T-1)-K]$. For robust estimates of the standard errors for the FE model, see Arellano (1987).

For the random effects model, $\mu_i \sim \text{IID}(0, \sigma_\mu^2)$, $\nu_{it} \sim \text{IID}(0, \sigma_\nu^2)$ and the μ_i's are independent of the ν_{it}'s. In addition, the X_{it}'s are independent of the μ_i and ν_{it}, for all i and t. The random effects model is an appropriate specification if we are drawing N individuals randomly from a large population, and we have no endogeneity between the regressors and the disturbances. For household panel studies, special attention is usually taken in the design of the panel to make it 'representative' of the population we are trying to make inferences about. In this case, N is usually large, and a fixed effects model would lead to an enormous loss of degrees of freedom. The individual effect is characterized as random, and inference pertains to the population from which this sample was randomly drawn. But what is the population in this case? Nerlove and Balestra (1992) emphasize Haavelmo's (1944) view that the population 'consists *not* of an infinity of individuals, in general, but of an infinity of *decisions*' that each individual might make. They argue that the fixed effects model may be more appropriate in cases where the population is sampled exhaustively (like data from geographic regions over time), whereas the random effects model is more consistent with Haavelmo's view given above. They argue that what differentiates individuals, who make the decisions with which we are concerned, is largely historical. Taking a leaf from Knight (1921), they argue that these inheritances from the past are material goods and appliances, knowledge and skill, and morale. In a dynamic context, this means that the primary reasons for heterogeneity among individuals is the different history each one has.

The random effects model implies a homoskedastic variance $\text{var}(u_{it}) = \sigma_\mu^2 + \sigma_\nu^2$ for all i and t, and an equi-correlated block-diagonal covariance matrix which exhibits serial correlation over time only between the disturbances of the same individual. In fact,

$$\text{cov}(u_{it}, u_{js}) = \sigma_\mu^2 + \sigma_\nu^2 \text{ for } i = j, t = s = \sigma_\mu^2 \text{ for } i = j, t \neq s$$

and zero otherwise. This also means that the correlation coefficient between u_{it} and u_{js} is

$$\rho = \text{correl}(u_{it}, u_{js}) = 1 \quad \text{for} \quad i = j, t = s$$
$$= \sigma_\mu^2 / \left(\sigma_\mu^2 + \sigma_\nu^2\right) \quad \text{for} \quad i = j, t \neq s$$

and zero otherwise. In this case, the BLUE of the regression coefficients is GLS which can be obtained from a least squares regression of $y_{it}^* = y_{it} - \theta \overline{y}_{i.}$ on $X_{it}^* = X_{it} - \theta \overline{X}_{i.}$ and a constant (see Fuller and Battese 1974). The GLS estimator of β for this random effects model will be denoted by $\widehat{\beta}_{RE}$. Here $\theta = 1 - (\sigma_\nu = \sigma_1)$ and $\sigma_1^2 = T\sigma_\mu^2 + \sigma_\nu^1$. Note that (i) if $\sigma_\mu^2 = 0$ then $\theta = 0$ and $\widehat{\beta}_{RE}$ reduces to $\widehat{\beta}_{OLS}$ since y_{it}^* reduces to y_{it}; (ii) if $T \to \infty$, then $\theta \to 1$ and $\widehat{\beta}_{RE}$ tends to $\widehat{\beta}_{FE}$ since y_{it}^* reduces to $\tilde{y}_{it}$. The variance components can be estimated from the between- and within-variation of the disturbances:

$$\sigma_1^2 = T \sum_{i=1}^{N} \widehat{\overline{u}}_{i.}^2 / (N - K - 1) \tag{7}$$

and

$$\widehat{\sigma}_1^2 = \frac{\sum_{i=1}^{N} \sum_{t=0}^{T} \tilde{u}_{it}^2}{[N(T-1) - K]} \tag{8}$$

where $\widehat{\overline{u}}_{i.}$ denotes the between-residuals from (4). Note that (7) is T times the s^2 of the between-regression obtained in (4). Also, $\sim \tilde{u}_{it}$ denotes the FE residuals from (5). So, (8) is the s^2 of the FE regression obtained in (5). Substituting these estimates for the variance components in θ and running y_{it}^*on X_{it}^* yields a feasible GLS or RE estimator suggested by Swamy and Arora (1972). For alternative estimators of the variance components, see Baltagi (2005). These are implemented using standard econometric software, including EViews, Stata, TSP, RATS and LIMDEP, to mention a few.

After this discussion of the fixed effects and the random effects models and the assumptions

underlying them, the reader is left with the daunting question: which to choose? This is not as easy a choice as it might seem. In fact, the fixed versus random effects issue has generated a hot debate in the biometrics and statistics literature, which has spilled over into the panel data econometrics literature. Economists cannot perform natural experiments of, say, the effect of fertilizer brand on crop yield controlling for the effect of land and other inputs. We have to deal with human subjects whose individual effects may be correlated with the regressors even when we randomly draw these individuals. Mundlak (1961) and Wallace and Hussain (1969) were early proponents of the fixed effects model, and Balestra and Nerlove (1966) were advocates of the random effects model. The modern econometric interpretation of the μ_i's is that they are random variables but in the RE model the $E(\mu_i = X_{it}) = 0$. This implies that the individual effects are uncorrelated with the regressors. This is a strong assumption given economists preoccupation with endogeneity issues. For example, in an earnings equation, μ_i may denote the unobservable ability of the individual and this may be correlated with the schooling variable included as a regressor. In this case, $E(\mu_i = X_{it}) \neq 0$ and the RE estimator $\widehat{\beta}_{RE}$ becomes biased and inconsistent for β. However, the within-transformation wipes out these μ_i's and leaves the FE estimator $\tilde{\beta}_{RE}$ unbiased and consistent for β. Hausman (1978) suggested comparing $\widehat{\beta}_{RE}$ and $\tilde{\beta}_{RE}$, both of which are consistent under the null hypothesis H_0; $E(\mu_i/X_{it}) = 0$. In this case, the contrast $\widehat{q} = \widehat{\beta}_{RE} - \tilde{\beta}_{RE}$ will have plim $\widehat{q} = 0$ under H_0. However, if H_0 is not true, plim $\widehat{q} \neq 0$ and the Hausman test statistic is given by

$$m = \widehat{q}'[\text{var}(\widehat{q})]^{-1}\widehat{q} \tag{9}$$

Under H_0 this is asymptotically distributed as χ^2_K where K denotes the dimension of slope vector β. For significant values of m, we reject the consistency of the RE estimator. Since $\widehat{\beta}_{RE}$ is the efficient estimator under the null hypothesis H_0, one can show that the cov$(\widehat{q}, \widehat{\beta}_{RE}) = 0$ and that the var $(\widehat{q})$ = var $(\widehat{\beta}_{FE})$- var$(\widehat{\beta}_{RE})$. This makes the computation of (9) simple. Nevertheless, Hausman (1978) suggested an alternative asymptotically equivalent test to (9) that can be obtained from the augmented regression

$$y^* = X^*\beta + \tilde{X}\gamma + w \tag{10}$$

where $y^*_{it} = y_{it} - \theta\overline{y}_i$, $X^*_{it} = X_{it} - \theta\overline{X}_i$; and $\tilde{X}_{it} = X_{it} - \overline{X}_i$. Hausman's test is now equivalent to testing whether $\gamma = 0$. This is a standard Wald test for the omission of the FE regressors $\tilde{X}$ from the RE regression. For an alternative variable addition test that produces a Hausman test which is robust to autocorrelation and heteroskedasticity of arbitrary form, see Arellano (1993).

Note that the FE model allows for endogeneity of the regressors and the individual effects, whereas the RE model does not. This is why the FE model is more popular among economists. Mundlak (Mundlak 1978) assumed that the individual effects are a linear function of the averages of *all* the explanatory variables across time, that is,

$$\mu_i = \overline{X}_i'\pi + \varepsilon_i \tag{11}$$

where $\varepsilon_i \sim \text{IIN}(0, \sigma^2_\varepsilon)$ and $\overline{X}_i'$ is $1 \times K$ vector of observations on the explanatory variables averaged over time. These effects are uncorrelated with the explanatory variables if and only if $\pi = 0$. In fact, a test for $\pi = 0$ yields the Hausman (1978) test based on the contrast between the FE and the between-estimators. Mundlak (1978) shows that GLS on (3) augmented with (11) yields $\tilde{\beta}_{FE}$. Only if $\pi = 0$ does it yield $\widehat{\beta}_{RE}$. This all-or-nothing choice of correlation between the individual effects and the regressors prompted Hausman and Taylor (1981) to suggest a model where *some* of the regressors are correlated with the individual effects. They proposed an instrumental variable estimator, denoted by HT, which uses both the between- and within-variation of the strictly exogenous variables as instruments. More specifically, the individual means of the strictly exogenous regressors are used as instruments for the time invariant regressors that are correlated with the individual effects (see Baltagi 2005, for more details).

F

The over-identification conditions are testable. In fact, this is a Hausman test based upon the contrast between the FE and the HT estimators.

Most applications in economics since the 1980s have made the choice between the RE and FE estimators based upon the standard Hausman test. If this standard Hausman test rejects the null hypothesis that the conditional mean of the disturbances given the regressors is zero, the applied researcher reports the FE estimator. Otherwise, the researcher reports the RE estimator. Unfortunately, applied researchers have interpreted a rejection as an adoption of the fixed effects model and non-rejection as an adoption of the random effects model. Chamberlain (1984) showed that the fixed effects model imposes testable restrictions on the parameters of the reduced form model and one should check the validity of these restrictions before adopting the fixed effects model (see also Angrist and Newey 1991). For the applied researcher, performing fixed effects and random effects and the associated Hausman test, it is important to carry this analysis a step further. Test the restrictions implied by the fixed effects model derived by Chamberlain (1984) before accepting the FE estimator and check whether a Hausman and Taylor (1981) specification might be a viable alternative.

See Also

- ▶ Artificial Regressions
- ▶ Dummy Variables
- ▶ Haavelmo, Trygve (1911–1999)
- ▶ Linear Models

Bibliography

Angrist, J.D., and W.K. Newey. 1991. Over-identification tests in earnings functions with fixed effects. *Journal of Business and Economic Statistics* 9: 317–323.

Arellano, M. 1987. Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics* 49: 431–434.

Arellano, M. 1993. On the testing of correlated effects with panel data. *Journal of Econometrics* 59: 87–97.

Balestra, P., and M. Nerlove. 1966. Pooling cross-section and time-series data in the estimation of a dynamic model: The demand for natural gas. *Econometrica* 34: 585–612.

Baltagi, B.H. 2005. *Econometric analysis of panel data*. Chichester: Wiley.

Chamberlain, G. 1984. Panel data. In *Handbook of econometrics*, ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Fuller, W.A., and G.E. Battese. 1974. Estimation of linear models with cross-error structure. *Journal of Econometrics* 2: 67–78.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12(Supplement): 1–118.

Hausman, J.A. 1978. Specification tests in econometrics. *Econometrica* 46: 1251–1271.

Hausman, J.A., and W.E. Taylor. 1981. Panel data and unobservable individual effects. *Econometrica* 49: 1377–1398.

Knight, F.H. 1921. *Risk, uncertainty and profit*. Boston: Houghton Mifflin.

Lancaster, T. 2000. The incidental parameter problem since 1948. *Journal of Econometrics* 95: 391–413.

Mundlak, Y. 1961. Empirical production function free of management bias. *Journal of Farm Economics* 43: 44–56.

Mundlak, Y. 1978. On the pooling of time series and cross-section data. *Econometrica* 46: 69–85.

Nerlove, M., and P. Balestra. 1992. Formulation and estimation of econometric models for panel data. In *The econometrics of panel data: Handbook of theory and applications*, ed. L. Matyas and P. Sevestre. Dordrecht: Kluwer.

Neyman, J., and E.L. Scott. 1948. Consistent estimation from partially consistent observations. *Econometrica* 16: 1–32.

Swamy, P.A.V.B., and S.S. Arora. 1972. The exact finite sample properties of the estimators of coefficients in the error components regression models. *Econometrica* 40: 261–275.

Wallace, T.D., and A. Hussain. 1969. The use of error components models in combining cross-section and time-series data. *Econometrica* 37: 55–72.

Fixed Exchange Rates

Peter M. Oppenheimer

An exchange rate is a price of one currency in terms of others. The existence of exchange rates derives from the fact that the world is divided into a large number of currency areas, mostly coterminous with nationstates, which trade with one

another and therefore exchange currencies at some point (or else confine their trade to barter or 'counter-trade'). The monetary authorities of a country, which regulate money supply and credit conditions, have by the same token a responsibility for the country's exchange rate. The precise significance of the exchange rate in relation to economic policy depends on how that responsibility is exercised; in particular on how far the authorities decide to 'fix' the rate, i.e. keep its movement within a narrow band of fluctuation (in the limit, zero) over a period of time.

There are two polar cases. At one extreme monetary authorities may commit themselves to holding the exchange rate fixed on a quasi-permanent basis. This was the case with adherents to the gold standard before 1914, who defined their currency units in terms of a physical quantity of gold which was not intended to be altered in ordinary circumstances (i.e. short of war or general political breakdown). The gold parity was underwritten by official readiness to buy and sell bullion at the declared price in terms of national currency. The currency exchange rate could then fluctuate in the market only within a narrow band around the parity, limited from above by the so-called gold-import point (at which it would be just profitable for gold traders to ship gold in from abroad for sale to the monetary authorities) and from below by the corresponding gold-export point.

At the other extreme is the case of a freely floating exchange rate. Here the authorities refrain not only from declaring any kind of exchange parity for the currency, but also from intervening in the currency market in order to stabilize or influence the rate. Their impact on the rate is then purely indirect (via the influence of monetary, fiscal and other policies on the behaviour of exchange-market participants), aside from any external transactions undertaken as part of the ordinary business of government, e.g. loans to foreign governments or expenditure on the diplomatic service.

In between the two extremes is a variety of possible exchange rate arrangements. Fixity of rates becomes a matter of degree. The International Monetary Fund (IMF) Articles of Agreement, adopted after the Bretton Woods Conference of 1944, required currencies to be given a par value in terms of gold (either directly or via the US dollar which itself was defined in terms of gold); but the par values could be altered in the event of 'fundamental disequilibrium' and thus came to be known as 'adjustable pegs'. Going down the spectrum, criteria for altering parities can be set so as to encourage more frequent and presumably smaller changes (as in the various types of 'sliding' or 'crawling' peg regimes), and the permitted margins of fluctuation around any given peg can be widened. If parities are abandoned, the authorities may still engage in extensive management of the floating rate through intervention in the currency market ('dirty floating'), as well as measures of monetary policy or exchange control.

The choice of exchange-rate arrangements for a single country is constrained by circumstances in the world at large and/or by the nature of the country's own economy. If, for example, major countries form a fixed-rate system, then an individual small country will have the choice of either participating in the system or remaining outside it and selecting its own exchange-rate regime. If, on the other hand, the major currencies are floating in relation to one another (like the US dollar, the yen and the Deutschemark after 1973), then there is no straightforward fixed-rate option for other countries. At best, they can peg their currencies to *one* of the majors, or they can stabilize the value of their own currency in terms of some 'basket', i.e. weighted average of foreign currencies.

The Price Level and Monetary Stability

Whether freely chosen or not, a country's exchange-rate regime affects, first, the dynamic relationship between its national price level and those of other countries, and secondly, the *modus operandi* and relative impact of monetary and fiscal policy instruments.

The more rigidly fixed a country's exchange rate, the greater is the weight of external influences in determining movements of its domestic price level. The channels through which these

external influences make themselves felt are varied and complex. They are relatively direct in the case of goods, services and factors of production traded internationally. To be sure, transport and transaction costs, product differentiation and other market imperfections prevent full compliance with the 'law of one price' even for the traded goods sector; but the sum total of such obstacles to full price equalization tends to be relatively constant over time, so that any significant change in the world price of a country's imports or exports is quickly passed through.

For the change in question to be, and to remain, purely monetary in nature, i.e. to have no impact on the level or composition of output and real incomes, three further conditions must be fulfilled. First, the global price shock must itself be purely monetary, i.e. must affect all traded-goods prices equiproportionally and leave the terms of trade unaltered. Secondly, the domestic economy must be characterized by widespread price flexibility, so that the price impulse is promptly transmitted to non-traded items, thus leaving domestic relative prices (of traded and non-traded goods) also unaltered. Thirdly, there must be appropriate adjustments in macro-economic, especially monetary, policy, in order to prevent either over-financing or under-financing of a given real product as the price level changes.

These conditions will seldom be met simultaneously. Monetary and real (output) disturbances are in practice intermingled. However, the conspicuous feature of fixed exchange rates in this domain is that they enforce, or presuppose, an approximately uniform system-wide inflation rate (as under the pre-1914 gold standard, or in the adjustable-peg period of 1950–1970). By contrast, floating rates permit wide divergences in national inflation rates, which are accommodated, and in part brought about, by exchange-rate movements (as was widely seen in the 1970s). Systemic inflation in the presence of fixed exchange rates will in practice always be low; otherwise the system would not command wide acceptance.

The combination of low inflation and a fixed or pegged exchange rate constitutes a virtual definition of monetary stability in an international system, and provides major real benefits by facilitating the efficient operation of the price system and the near-optimal use of money in exchange. Nonetheless, depending on the precise constitution of a fixed-rate system (i.e. whether rates are meant to be totally rigid; or if not, in what conditions and by how much they may be altered), countries may opt out or may be forced out for either of two reasons. They may find the international inflation rate unpalatable (e.g. a rate of three per cent per annum is probably acceptable to many countries but distastefully high to a few), and see insufficient compensating attractions in exchange-rate fixity as such. Alternatively, they may find the international inflation rate unattainably low, at any rate without incurring, or appearing to incur, unacceptable (even if temporary) costs. The costs comprise lost output and employment, or social disruptions over price/wage issues such as subsidies or trade union reform.

Monetary and Fiscal Policy

A pegged exchange rate calls for a certain pattern of macro-economic management by national authorities. Monetary policy, especially changes in interest rates, can play a leading role in influencing aggregate private spending only if there are narrow limits to the international mobility of funds. Otherwise, the main impact of monetary measures, at least up to the medium term, is upon the disposition of internationally mobile stocks of capital, and hence upon the financial underlay to a given volume and value of national expenditures, rather than the expenditure volume itself. Monetary tightening pulls in funds from abroad; monetary easing pushes funds out (unless the respective tightening and easing is simultaneously matched by other countries). In the limit, national interest rates are determined wholly by the international capital market and its assessment of the individual country's credit rating, rather than by national preferences or policy. By the same token, fiscal policy (government expenditure, taxation and borrowing) then has a relatively great impact on national expenditure, output and the external current-account (export/import) balance.

The division of function between policy variables is quite different with freely floating exchange rates. Here the international mobility of capital (without which floating is not feasible) means that monetary measures, instead of affecting the level of external reserves, alter the exchange rate and thus the domestic price of traded goods. This in turn, depending on circumstances as before, will affect the price of non-traded goods and/or the level and composition of output. Monetary expansion, for instance, depreciates the exchange rate and raises the domestic price of traded goods, stimulating the economy and generating some combination of higher output and higher prices. Pure fiscal policy, on the other hand, is generally less effective than before, because higher (lower) public-sector borrowing demands lead promptly to a higher (lower) exchange rate, which tends to offset the aggregate expenditure impact of the fiscal change. Only in the special case of balanced-budget fiscal policy may an equiproportionate change in government outlays and receipts affect aggregate demand even in a floating-rate regime with perfect world capital markets (McKinnon and Oates 1966).

The contrast between the fixed and floating rate cases is most complete for a 'small' country whose behaviour has no significant impact on global economic variables. In a 'large' country monetary tightening will influence credit conditions worldwide under both fixed and floating rates, while fiscal policy will have an impact on aggregate world expenditure under either exchange-rate regime.

The World Monetary System

Exchange-rate arrangements are the most important constituent of the (market-economy) world monetary system. The other principal constituents are international reserve assets and arrangements for co-operation among sovereign monetary authorities and (where appropriate) international bodies such as the IMF. Global exchange-rate arrangements are determined by the small number of major countries which at any one time constitute the core of the international economy. After 1973 the world was perceived to have abandoned the pegged-rate system in favour of floating rates, even though the vast majority of the world's 100-plus currencies remained pegged to some major currency or basket of currencies. The crucial change lay in the fact that the US dollar, the Deutschemark and the yen were now in a floating relationship to one another. In addition, a few currencies of secondary importance, such as the pound sterling and the Swiss franc, were likewise floating.

The principal focus of the story is the dollar, as the system's principal reserve currency and the currency in terms of which virtually all countries had maintained pegged exchange rates over the preceding quarter-century. The key question is why the German and Japanese authorities did not re-establish a pegged-rate relationship with the dollar, despite great concern at times over the way in which floating rates were moving. Indeed, in 1978 the German government specifically took the initiative to create within Europe a stronger bloc of mutually pegged exchange rates (the European Monetary System) as a counterweight to an unstable and at that time undervalued dollar. Evidently, pegging to the dollar was seen as courting greater risks to financial stability than other courses of action. Such risks must be rooted in the presumed determinants of US financial policy, and specifically in the belief that, if other major countries commit themselves to maintaining exchange-rate pegs vis-à-vis the dollar, the US authorities for their part will not give adequate weight to the external repercussions of their policy unless they too are committed to defending an exchange-rate peg and reserve position of their own. A pure 'dollar standard' has not been an acceptable basis for a world-wide system of pegged rates, because it would leave the United States insufficiently subject to balance-of-payments discipline.

The problem of imposing payments discipline on the centre country (or countries) of a fixed-rate system has historically been solved (or avoided) in only one way, namely by pegging that country's currency and hence the system as a whole to an 'outside' commodity asset, most successfully to gold. The market for this commodity then serves

as the vehicle for reconciling the competing responsibilities and preferences of the sovereign governments which make up the international system.

The theory of the pre-1914 gold standard was that movement of gold reserves determined changes in national money stocks and hence, with a given structure of domestic payments, in money national incomes. Price and wage flexibility was relied upon to assure full employment of available productive resources and, in the process, to reconcile the resulting real national incomes with their current money values as determined by the monetary mechanism.

Further implications followed. The distribution of global increments in the stock of monetary gold (equal in any period to the excess of current mine production over net private offtake for industry, hoarding, etc.) was governed by relative growth rates of real GNP. Fast growth of an economy tended to produce a relative lowering of its price level, which tendency would be checked and the price level kept in line by relatively fast growth of its gold reserves and money supply. Finally, if global economic growth was faster (slower) than current growth of money stocks, there would be downward (upward) pressure on the world price level; with the price of gold alone fixed in money terms, this meant a rise (fall) in the real price of gold, which would sooner or later augment (diminish) the net inflow of gold to the monetary system, thereby tending to halt or reverse the original movement in world price levels.

The operation of the gold standard in practice corresponded only very partially to the theoretical model. For instance, growth of monetary gold stocks was reconciled with faster growth of national outputs less by downward pressure on price levels than by increased concentration of monetary gold at central bank reserves and a shrinkage in gold's share of money aggregates (Triffin 1964). However, national monetary policies were governed to a large degree by balance-of-payments considerations, and a broad measure of global price stability was maintained.

The adjustable-peg system of Bretton Woods (devised chiefly by J.M. Keynes and H.D. White) was a type of gold-exchange standard, but one which ultimately subordinated changes in monetary gold stocks to the growth of money incomes rather than the other way round. This intended reversal of gold-standard relationships stemmed from the Keynesian assumptions that maintenance of full employment was a government responsibility which could not in general be delegated to market forces, and that money wages and prices were inclined to be inflexible, especially downwards; hence national authorities must be free to arrange whatever level of national purchasing power they judged appropriate for maintaining high employment and avoiding inflation. Situations might arise in which one or more countries could not achieve this overriding objective at the previously declared exchange-rate pegs ('par values') without recourse to (additional) administrative restrictions on trade and current payments. In such a situation (the 'fundamental disequilibrium' of IMF terminology) a par value could be adjusted – downwards to reduce the home country's wage level in international terms, thus boosting its competitiveness; or upwards to increase the wage level in international terms, thus fending off excessive reserve gains and inflation.

Modest reserve gains, however, were viewed as desirable, and certainly as acceptable, by many countries, particularly in a period of rapid economic expansion like the 1950s and 1960s. Equilibrium of the system as a whole therefore required a certain growth of global exchange reserves to avoid a competitive scramble among countries for balance-of-payments surpluses. The annual inflow of new monetary gold after 1945 was at no time sufficient for this purpose. The gap was filled, at first deliberately and then involuntarily, by the United States, which ran an overall deficit on its balance of payments, thereby acting as a net supplier of reserves to other countries. The immediate supply took the form of dollars, which then constituted a potential claim on the US gold stock and were in part exchanged for gold by foreign monetary authorities.

Triffin (1960) first emphasized that this process was weakening the external liquidity position of the United States and could not continue indefinitely without calling into question the gold

convertibility of the dollar at its declared par value of 0.888671 grammes of gold fine or $35 per ounce of gold. Contrary, however, to what Triffin implied, the United States could not put an end to its deficit without first altering (or abandoning) its par value. By standing ready to sell gold to foreign monetary authorities at $35 an ounce, the US Treasury was acting in effect as buffer-stock manager for an under-priced commodity – a commitment which could have only one outcome. Perception of the point was paradoxically hampered by the fact that the dollar was until near the end of the 1960s scarcely overvalued against other major currencies. The pressure on the US balance of payments to act as a net source of reserves to the outside world stemmed from the dollar's overvaluation in common with all other currencies vis-à-vis gold (Gilbert 1968, 1980).

The IMF Articles had envisaged such a possibility. Not only did they give the United States exactly the same scope to alter its par value as any other country; in addition, they provided for 'a uniform change in all par values', i.e. a general rise in the price of gold, in order to relieve a system-wide shortage of reserves or reserve increments. The US authorities declined to avail themselves of this measure, viewing or professing to view it as unlikely to promote payments equilibrium and therefore as an unwarranted blow to the prestige of the dollar. By 1970 US gold reserves had declined from their post-World War II peak of $22 billion to little more than $10 billion, while US liquid external liabilities had risen from negligible amounts to over $20 billion. The dollar's gold convertibility was formally abrogated on 15 August 1971 and the attempt to maintain a pegged-rate system on the basis of an inconvertible dollar foundered in March 1973.

Many observers have been reluctant to accept that the demise of the pegged-rate system was due to the US refusal to increase the dollar price of gold. Instead they have claimed, on the one hand, that the Bretton Woods system would in any event have been swept away by the inflation and balance-of-payments problems of the 1970s (an unconvincing line of argument, not least because the world inflation of the 1970s was itself in large measure caused by the financial turmoil in which the pegged-rate system collapsed); and on the other hand, that a fiduciary asset such as IMF Special Drawing Rights could have replaced gold (and could still do so) at the base of a pegged-rate system, but for the fact that the vulnerability of adjustable pegs to speculative attack renders them unviable anyhow in the face of free international capital movements.

Neither leg of the latter argument is persuasive. Gold was able to function as the basis of an adjustable peg system because its availability for this purpose is regulated with the help of market forces and without the need for detailed and continuous agreement on reserve creation and exchange-rate policy among sovereign governments. Specifically, the Bretton Woods System incorporated a strong and direct link between the exchange-rate policy and the international liquidity position of the United States: a reduction in the dollar's par value could always be made large enough to produce a decisive impact on US reserves. A fixed-rate system based on a fiduciary asset such as SDRs would lack this feature, and would therefore be only a special form of currency standard, like the abortive dollar standard of 1971–73.

Currency speculation, as distinct from politically motivated capital flight, does not initiate balance-of-payments problems. Rather, it emerges as an aggravating factor when there is an evident underlying disequilibrium which the authorities are slow to tackle and which therefore presents speculators with the prospect of easy gains. Variation in the method of altering an individual par value (e.g. temporary floating, or small changes of greater frequency) may in some circumstances be a useful means of containing or discouraging speculation. Such devices, however, were quite irrelevant to the disequilibrium and breakdown of Bretton Woods, since the United States was unwilling to alter the dollar's par value by any method, and without such alteration the system could not be brought to equilibrium.

See Also

► Crawling peg
► Flexible exchange rates

- ▶ International finance
- ▶ International monetary policy

Bibliography

Dornbusch, R. 1980. *Open economy macro-economics*. New York: Basic Books.

Eichengreen, B. (ed.). 1985. *The gold standard in theory and history*. New York/London: Methuen.

Fleming, J.M. 1962. Domestic financial policies under fixed and under floating exchange rates. *IMF Staff Papers* 9(November): 369–379.

Gilbert, M. 1968. *The gold/dollar system: Conditions of equilibrium and the price of gold*, Princeton Essays in International Finance No. 70. Princeton: Princeton University Press.

Gilbert, M. 1980. *Quest for world monetary order*. New York: Wiley.

McKinnon, R., and W.E. Oates. 1966. *The implications of international economic integration for monetary, fiscal and exchange-rate policy*, Princeton Studies in International Finance No. 16. Princeton: Princeton University Press.

Meade, J.E. 1951. *The theory of international economic policy: vol. I, The balance of payments*. London: Oxford University Press.

Mundell, R.A. 1968. *International economics*. New York: Macmillan.

Triffin, R. 1960. *Gold and the dollar crisis*. New Haven: Yale University Press.

Triffin, R. 1964. *The evolution of the international monetary system: Historical Reappraisal and Future Perspectives*, Princeton Studies in International Finance No. 12. Princeton: Princeton University Press.

Fixed Factors

Walter Y. Oi

Abstract

Small firms invest relatively less in custom-made machines and specifically trained employees. The overhead costs of fixed-capital assets are relatively larger for big firms that engage in the volume production of standardized products. Large firms also incur higher fixed employment costs to recruit and train a specialized workforce. Workers in large firms are paid higher wages designed to reduce labour turnover rates. These phenomena could not be explained without a formal analysis of fixed and quasi-fixed factors. A continuum of degrees of fixity makes for a richer theory of factor markets than a dichotomy of fixed versus variable factors.

Keywords

Amortization; Barriers to entry; Clark, J. M.; Elasticity of substitution; Firm size; Firm-specific factors; Fixed factors; Human capital; Implicit contracts; Labour as a quasifixed factor; Labour market search; Labour markets contracts; Monitoring costs; Overhead costs; Rationing; Shadow price; Specialization; Substitutes and complements; Training; Wage differentials

JEL Classifications

D5

In moving from one market equilibrium to another, a firm may choose to hold fixed the rate of employment of one or more factors of production. The presence of fixed factors and their associated overhead costs will affect the firm's responses to changing market conditions. The residually determined quasi-rents which constitute the returns to the fixed factors must, in the long run, cover their overhead costs; otherwise, the inputs of fixed factors have to be contracted. The importance of fixed factors and overhead costs, which varies across firms and industries, was analysed by J.M. Clark (1923), who emphasized the first of the following three questions: (1) How do fixed factors affect the behaviour of prices, outputs and inputs of variable factors? (2) What determines whether a factor of production will be fixed or variable? (3) How do the fixed employment costs of quasi-fixed labour inputs affect contractual arrangements in labour markets?

In the short run, certain paths of adjustment are barred to the firm. The usual assumption is that the input of one or more factors is fixed. Total unit costs, which include the outlays for fixed factors, lie above average variable costs so that price, in

the short run, can remain well below the minimum long-run average cost. If fixed costs in an industry are high, they can pose a barrier to entry of new firms and could result in wide short-run fluctuations in price. Further, the upper-bound constraint on inputs of fixed factors affects the firm's demand for the remaining variable inputs in a manner analogous to the theory of rationing of consumer goods analysed by E. Rothbarth (1941). An increase in the demand for the final product raises the shadow price of the fixed factor, which increases the demand for variable factors that are substitutes for the fixed factor and decreases the demand for complementary variable factors. This result could explain the greater cyclical volatility in the demand for unskilled labour relative to skilled labour if unskilled labour is a closer substitute for the fixed factor, capital. Moreover, the smaller the elasticity of substitution of labour for capital, the steeper is the slope of the marginal cost curve, implying larger cyclical swings in product prices.

A firm will fix the input rate of a factor if (*a*) the factor is specific to the firm in the sense that employment in this firm constitutes its highest valued use, or (*b*) reallocation to some higher-valued use is precluded by some contractual agreement or by a prohibitively high transaction cost. In the former case, equipment, buildings and even labour can be specialized to fit into a firm's idiosyncratic production methods. The internal values of such specialized resources are likely to exceed their external values to outside users. These resources are more likely to be owned (rather than hired or leased), because of their specificity. Long-term contracts that account for some fixed factors occur where there are gains from risk-sharing or high costs of transferring resources to other firms.

A richer theory of factor markets can be developed if the dichotomy of fixed versus variable factors is replaced by a continuum of degrees of fixity. The discipline of labour economics has now accepted the principle that labour is a quasi-fixed factor. The cost of hiring and training workers constitutes the fixed component of the full cost of labour, while the variable component is the wage paid to the employee.

In long-run equilibrium, the expected marginal value product which depends on the expected product price P^* and labour's marginal physical product f_N, is equated to the full labour cost:

$$P^* f_N = W + q, \left[q = \left(\frac{F}{r} \right) \left(1 - e^{-rT} \right) \right]$$

where W is the wage, and q is the periodic rent that amortizes the fixed employment cost F at a discount rate r over the worker's expected period of employment T. The gap between the wage and labour's marginal value product will be relatively larger, the higher is the degree of fixity which can be measured by $f = q/W + q$.

The cyclical behaviour of the labour market is characterized by an uneven incidence of unemployment, a compression of occupational wage differences in the upswing, persistent differences in labour turnover rates, hiring/firing practices that smack of discrimination. The quasi-fixity of labour goes a long way in explaining these phenomena. In the downswing, the product price falls below its long-run level P^*. If labour is a completely variable input, meaning that $q = F = 0$, its marginal value product Pf_N will be equated to the wage in each period. Hence, when P falls, the demand for this grade of labour is contracted until f_N climbs to restore equilibrium in both factor and product markets. However, if labour is a quasi-fixed factor, the periodic amortization of the fixed cost drives a wedge between the wage and marginal value product. For a small decline in product price, the firm will not contract the demand for a quasi-fixed grade of labour as long as its short run MVP exceeds the wage, which is the variable cost of labour; that is, if $Pf_N > W$ even though $Pf_N < (W + q)$, the input of this grade of labour will not be reduced in the downswing. There is, for each quasi-fixed factor, a trigger price P_i at which the firm will choose to reduce employment. The trigger price which induces a decline in factor demand will be lower for factors with higher degrees of fixity. In the early stages of a downturn, labour with low degrees of fixity will become unemployed, while other workers will be retained until the drop in product price P is driven below P_T. At the trough of a cycle,

most grades of labour satisfy a short-run equilibrium condition where labour's MVP is equated to its variable cost, $Pf_N = W$. As P rises in the recovery, a firm will increase its demand for a quasi-fixed factor if the price rise is such that labour's MVP exceeds its *full cost*; that is, employment is expanded if and only if $Pf_N > (W + q)$. In the upturn, the rightward shift in factor demand will be greater for factors with lower degrees of fixity. Employment will be more stable, and the incidence of unemployment will be lower for those workers in occupations with higher degrees of fixity.

Some firms find that it is profitable to incur the fixed employment costs of assembling a firm-specific workforce. Recruiting is the means by which an employer identifies more productive individuals and ascertains whether an applicant will meet prescribed hiring standards. Recruitment for high-wage positions usually entails higher costs because of the variability of individual productivities. Employers who have well-defined internal labour markets and who organize production around teams also incur higher recruiting costs. In an internal labour market, workers are hired at a limited number of ports of entry and are typically given on-the-job training to adapt them to the firm's idiosyncratic production methods. Larger investments in firm-specific human capital are indicative of the greater specialization of the labour input. Firm-specific training is less profitable when labour turnover rates are high due either to the high separation propensities of workers or the low survival odds of firms. Smaller firms spend less on recruiting and appear to invest less in formal training. The estimates reported by Oi (1962) and Parsons (1972) reveal that employers incurred substantially higher fixed employment costs for workers in higher skill levels. The degree of fixity, $f = q/(W + q)$, is positively related to the wage rate W, and this relation allows us to test the implications of a theory of labour as a quasi-fixed factor. Employees in high-wage occupations experience greater employment stability over the cycle. Occupational wage differentials widen in the downswing and narrow in the upswing. Labour turnover rates are lower, and recruiting costs are higher in large firms whose workforces exhibit a higher degree of fixity.

The persistence of unemployment and the failure of wages to clear labour markets call for an explanation. Some unemployed workers are in a state of pseudoidleness while they look for work: 'When actively searching for work, the situation is that he is really investing in himself by working on his own account without immediate remuneration. He is prospecting' (Hutt 1977, p. 83). The time and money spent by new entrants and disemployed workers in their search for suitable job matches constitute a fixed cost which has to be recovered over the course of the employment relation. Each job is, in a very real sense, specialized to the worker–firm attachment. In a search model, unemployment can be efficient in two senses. First, it may be the least-cost means of finding a durable job. Second, a worker on a temporary layoff may stay in a state of availability awaiting recall rather than seeking work. Labour turnover is costly, both to the employer for whom labour is a quasi-fixed factor due to the fixed investments in hiring and training, as well as to the employee for whom this job is specific due to the fixed costs of search. Both parties have incentives to form an implicit contract that can raise the returns to these fixed employment costs by lengthening the expected period of employment.

Long-term employment contracts could be the result of risk-averse workers seeking job security. An employer can reduce his full labour costs by providing a tacit agreement in which the risks of income variability are shared. Such long-term agreements end up increasing the fixity of labour. Implicit, long-term contracts may also result from an employer's desire to discourage shirking and dishonesty. Firms will incur monitoring and enforcement costs to deter dysfunctional behaviour and malfeasance. These enforcement costs can be reduced by designing compensation packages which reward workers with separation pay and pensions if they perform in accordance with prescribed work standards. Stable and durable employment relations make sense only when there are fixed costs of forging

and maintaining specific jobs defined by worker–firm attachments.

When physical or human capital is specialized to a firm, it must capture any quasi-rents that it can because the fixed investments in these specialized resources cannot be reallocated to some alternative use. Fixed, firm-specific factors only make sense in a world of heterogeneous firms. In Oi (1983) *I* advanced the thesis that firm-specific capital was systematically related to firm size. Small firms with low survival odds do not invest in custom-made machines and specifically trained employees. They are more likely to purchase used assets and to hire inexperienced workers with general human capital. The overhead costs of fixed-capital assets are relatively larger for big firms that engage in the volume production of standardized products. Large firms also incur higher fixed employment costs to recruit and train a specialized workforce. Workers in large firms are paid higher wages and are provided with employee compensation packages that are designed to reduce labour turnover rates. These phenomena could not be explained without a formal analysis of fixed and quasi-fixed factors.

See Also

► Rent

Bibliography

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

Hutt, W.H. 1977. *The theory of idle resources*. Indianapolis: Liberty Press.

Oi, W.Y. 1962. Labor as a quasi-fixed factor. *Journal of Political Economy* 70: 538–555.

Oi, W.Y. 1983. The fixed employment costs of specialized labor. In *The measurement of labor costs*, ed. J.E. Triplett. Chicago: University of Chicago Press.

Parsons, D.O. 1972. Specific human capital: An application to quit rates and layoff rates. *Journal of Political Economy* 80: 1120–1143.

Rothbarth, E. 1941. The measurement of changes in real income under conditions of rationing. *Review of Economic Studies* 8: 100–107.

Fixed Point Theorems

Andrew McLennan

Abstract

This article gives statements of the Tarski fixed point theorem and the main versions of the topological fixed point principle that have been applied in economic theory. Pointers are given to literature concerned with proofs of Brouwer's theorem, and with algorithms for computing approximate fixed points. The topological results are all consequences of a slightly weakened version of the Eilenberg and Montgomery (*American Journal of Mathematics* 68: 214–222, 1946) fixed point theorem. The axiomatic characterization of the Leray–Schauder fixed point index (which is even more powerful) is also stated, and its application to issues concerning robustness of sets of equilibria is explained.

Keywords

Absolute neighbourhood retract; Algebraic topology; Brouwer's fixed point th; Contraction mapping th; Convexity; Cooperative game theory; Cooperative game theory (core); Debreu–Gale–Kuhn–Nikaido lemma; Eilenberg–Montgomery th; Essential sets of fixed points; Excess demand; Existence of equilibrium; Fixed point property; Fixed point theorems; Homotopy methods; Hopf's th; Kakutani's th; Kinoshita's th; K–K–M–S th; Lefschetz fixed point th; Leray–Schauder fixed point index; Nash equilibrium; Perfect equilibrium; Sard's th; Scarf algorithm; Schauder fixed point th; Sperner's lemma; Strategic stability; Tarski's fixed point th

JEL Classifications

D5

The Brouwer (1910) fixed point theorem and its descendants are key mathematical results underlying the foundations of economic theory.

Let f: $X \to X$ be a function from a space to itself. A *fixed point* of f is a point $x^* \in X$ that is mapped to itself by f : $f(x^*) = x^*$. A *fixed point th* is a result asserting that, under some hypotheses, the set of fixed points of f is nonempty. A simple example with many applications is:

Theorem 1 (Contraction Mapping Th) *If the metric space (X, d) is complete (recall that this means that every Cauchy sequence is convergent) and there is a number* $c \in (0, 1)$ *such that* $d(f(x), f(x')) \leq cd(x, x')$ *for all* $x, x' \in X$, *then f has a unique fixed point.*

Another example illustrating the importance of the general notion of completeness, but otherwise based on quite different principles, is:

Theorem 2 (Tarski's (1955) Fixed Point Theorem) *Let* $(X, \leq)$ *be a complete lattice:* $\leq$ *is a partial ordering of X and every subset of X has a greatest lower bound and a least upper bound. If* f: $X \to X$ *is monotone – that is,* $f(x) \leq f(x')$ *whenever* $x \leq x'$ *– then there are fixed points* $\underline{u}, \overline{u} \in X$ *such that* $\underline{u} \leq x$ *whenever* $x \leq f(x)$ *and* $x \leq \overline{u}$ *whenever* $f(x) \leq x$.

This result is foundational for the theory of strategic complementarities – for example, Milgrom and Shannon (1994), Echenique (2005) – and has been applied to growth theory by Hopenhayn and Prescott (1992).

The rest of our discussion is devoted to results related to Brouwer's fixed point theorem. A topological space has the *fixed point property* if every continuous map from the space to itself has a fixed point. Brouwer's theorem states that a nonempty compact convex subset of a Euclidean space has the fixed point property. This celebrated result underlies many of the advanced results of topology, and was a pivotal event in the development of algebraic topology, which has influenced many areas of mathematics. In the half-century following Brouwer's paper the theory of fixed points was extended in various directions, yielding several generalizations of Brouwer's result that are themselves famous theorems. Early in the post-war period fixed point theorems were used by Arrow and Debreu (1954), McKenzie (1959), Nash (1950, 1951), and Debreu (1952) to prove the fundamental equilibrium existence results of theoretical economics: every economy with finitely many goods and agents has a competitive equilibrium; every finite normal form game has a Nash equilibrium. Fixed point theory continues to play an important role in the extensive body of research that grew out of these fundamental discoveries.

Useful books devoted to fixed point theory include Border (1985), which emphasizes results used in economic theory, Brown (1971), which develops the theory of the fixed point index using the methods of algebraic topology, and Dugundji and Granas (2003), which comprehensively surveys the topic from the point of view of applications to analysis and topology. The latter book features extensive historical information concerning the development, and the developers, of the subject.

Proofs and Algorithms

Since Brouwer's theorem is a breakthrough result, one should expect proofs to reveal deep mathematical principles, and in fact Brouwer's work was a major stimulus to the development of the subject that is now known as algebraic topology. Eventually Sperner (1928) distilled a relatively simple combinatoric argument out of the topological ferment of that era. Although this argument is the most popular in graduate education in economics, in the author's opinion the exposition in Milnor (1965) of an argument due to Hirsch is worth whatever additional effort it entails, because the student also learns Sard's theorem, which is another fundamental result of the 20th century with important applications in economic theory. Although the substance of the argument in Milnor (1978) appears to be less useful, its brevity and elementary character are stunning. The proof of McLennan and Tourky (2005) is also relatively simple, and displays how Kakutani's theorem

follows easily from the existence of Nash equilibrium for a special class of two-person games, which is one of the simplest manifestations of the fixed point principle.

Computation of approximate fixed points has many applications in economics and other fields, and is an important topic of research. Iteration of a function is guaranteed to work only when the function is a contraction, as in Theorem 1, but this method is often practical for functions that do not satisfy this condition. Other methods are derived from proofs of Brouwer's theorem. The method pioneered by Scarf (1973; Doup 1988) is a method of moving through the simplices of a simplicial subdivision of the simplex. It is justified by a refinement of the proof of Sperner's lemma. The proof derived from Sard's theorem points towards homotopy methods, which have a huge literature (Garcia and Zangwill 1981; Algower and Georg 1990). The proof in McLennan and Tourky (2005) also points towards algorithms in which the equilibria of certain two-person games give rise to approximate fixed points.

Variants

We will give statements of the main forms in which the fixed point principle is applied in economic theory. Let X and Y be metric spaces. A *correspondence* $F : X \to Y$ assigns a nonempty $F(x) \subset Y$ to each $x \in X$. When $Y = X$, a point x^* is said to be a *fixed point* if $x^* \in F(x^*)$. If P is any property of sets, then F is *P valued* if each image $F(x)$ has property P. It is *upper semicontinuous* (u.s.c.) if it is compact valued and, for each $x \in X$ and each neighborhood V of $F(x)$, there is a neighborhood U of x such that $F(x') \subset V$ for all $x' \in U$. It is not hard to show that if Y is compact, then F is u.s.c. if and only if its graph

$$Gr(F) = \{(x, y) \in X \times Y : y \in F(x)\}$$

is closed. We think of a function as a singleton-valued correspondence, in which case upper semicontinuity coincides with the usual notion of continuity.

Economic models frequently give rise to sets of optimal individual choices that are convex, but may have more than one element. For this reason the most prominent fixed point theorem in economic applications is:

Theorem 3 (Kakutani 1941) *If X is a nonempty compact convex subset of a Euclidean space and $F : X \to X$ is a u.s.c. convex valued correspondence, then F has a fixed point.*

The following variant is tailored for applications in general equilibrium theory, where one is searching for a price vector that equates supply and demand in all markets.

Theorem 4 (Debreu–Gale–Kuhn–Nikaido Lemma) *Let*

$$\Delta := \left\{ p \in \mathbb{R}^n_+ : \sum_{j=1}^{n} p_i = 1 \right\}$$

be the $n - 1$ dimensional simplex. If $Z : \Delta \to \mathbb{R}^n$ is a u.s.c.c.v. correspondence satisfying $p \cdot z = 0$ for all $p \in \Delta$ and all $z \in Z(p)$, then there is a $p^* \in \Delta$ and $z^* \in Z(p^*)$ such that $z^* \leq 0$.

The following result of Shapley (1973a, b; see also Herings 1997, and references cited therein) generalizes the famous K–K–M theorem of Knaster et al. (1929). It has important applications to the theory of the core and other aspects of cooperative game theory and general equilibrium theory.

Theorem 5 (K–K–M–S Th) *Let $\mathcal{N} = 2^{\{1,\ldots,n\}} / \emptyset$, and for $\mathcal{S} \in \mathcal{N}$ let $\Delta^S := \{x \in \Delta : x_i = 0$ for all $i \notin S\}$. If $\{C^S\}_{\mathcal{S} \in \mathcal{N}}$ is a collection of closed sets such that $\Delta^T \subset \cup_{S \subset T} C^S$ for all $T \in \mathcal{N}$, then there is $\mathcal{B} \subset \mathcal{N}$ and numbers $\lambda_S \geq 0$ for $S \in \mathcal{B}$ such that $\Sigma_{i \in S \in \mathcal{B}} \lambda_S = 1$ for all $i = 1, \ldots, n$, (such a $\mathcal{B}$ is called a balanced collection) and $\cap_{S \in \mathcal{B}} C^S \neq \emptyset$.*

The original K–K–M theorem is the special case in which $C^S = \emptyset$ whenever S has more than one element. That is, $C_1 \cap \ldots \cap C_n \neq \emptyset$ whenever $C_1, \ldots, C_n \subset \Delta$ are closed sets satisfying $\Delta^T \subset \cup_{i \in T} C_i$ for all $\mathcal{T} \in \mathcal{N}$.

Generalizations

During the first half of the 20th century there emerged a sequence of increasingly general versions of Brouwer's theorem. Let X and X' be metric spaces, and let $\varphi : X \to X'$ be a homeomorphism. A point $x^* \in X$ is a fixed point of a continuous function $f : X \to X$ if and only if $\phi(x^*)$ is a fixed point of $\varphi o f o \varphi^{-1}$, so the fixed point property is invariant under homeomorphism. Compactness and continuity are invariant properties, but the assumptions of convexity and finite dimensionality in Brouwer's theorem seem too strong, as does the assumption of convex valuedness in Kakutani's theorem. One is led to search for weaker, topological assumptions that imply the fixed point property.

Let Y be another metric space. A continuous function

$$h : X \times [0, 1] \to Y$$

is called a *homotopy*. For each $0 \leq t \leq 1$ let $h_t = h(\cdot, t) : X \to Y$. We think of 'continuously deforming' h_0 into h_1, with the variable t representing time, and we say that h_0 and h_1 are *homotopic*. The space X is *contractible* if the identity function on X is homotopic to a constant function. If X is convex, then for any $x_0 \in X$ the function

$$h(x, \ t) = x_0 + (1 - t)(x - x_0)$$

is such a homotopy, so convex sets are contractible. It was conjectured that nonempty compact contractible sets have the fixed point property, but eventually counterexs were discovered by Kinoshita (1953) and others.

A *retraction* of X onto a subset A is a continuous function $r : X \to A$ whose set of fixed points is A, so that $r(a) = a$ for all $a \in A$. In this circumstance we say that A is a *retract* of X One point of interest is that if X has the fixed point property, then so does A: if $g : A \to A$ is continuous, then $go\ r : X \to A \subset X$ has a fixed point x^*, and $x^* = g(r(x^*)) = g(x^*)$ because x^* must be in A.

The subspace A is a *neighbourhood retract* if there is an open $U \supset A$ and a retraction $r : U \to A$. A continuous function $e : X \to Y$ is an *embedding* if it is injective and $e^{-1} : e(X) \to X$ is continuous, that is, e is a homeomorphism onto its image. A metric space X is an *absolute neighbourhood retract* (ANR) if $e(X)$ is a neighbourhood retract whenever $e : X \to Y$ is an embedding of X in a metric space Y. The class of ANRs is large, encompassing many important types of spaces such as manifolds, simplicial complexes, and convex sets, and there is an extensive theory (for example, Borsuk 1967) that cannot be described here. One may think of an ANR as a space that has bounded complexity, in a certain sense, in a neighbourhood of each of its points. (An example of a space that is *not* an ANR is the union X of the unit circle centred at the origin in $\mathbb{R}^2$ and the set $\{(1 - \theta^{-1})(\cos\theta, \sin\theta : 1 \leq \theta < \infty\}$. If X was an ANR, then there would exist a retraction of a neighbourhood $U \subset \mathbb{R}^2$ onto X, and the retraction would take small connected neighbourhoods of (1, 0) in U to small connected neighbourhoods of (1, 0) in X, but small neighbourhoods of (1, 0) in X are disconnected.)

Eilenberg and Montgomery (1946) gave a fully satisfactory generalization of Brouwer's theorem: F has a fixed point whenever X is a nonempty compact acyclic ANR and $F : X \to X$ is a u.s.c. acyclic valued correspondence. Acyclicity is a concept from algebraic topology that cannot be defined here; the important point for us is that contractible sets are acyclic, and that the loss of generality in passing from acyclicity to contractibility is of slight concern in economic theory.

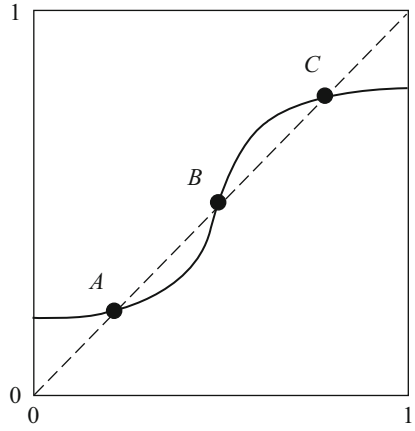
Contractible valued correspondences that are not convex valued appear in McLennan (1989a) and Reny (2005). There are many applications in economics of the special case of the Eilenberg–Montgomery theorem in which X is convex (but possibly infinite dimensional) and F is convex valued, for which relatively simple and direct proofs were given by Fan (1952) and Glicksberg (1952). In turn this result is more general than both Kakutani's theorem and the well known Schauder (1930) fixed point theorem.

The Leray–Schauder Fixed Point Index

Consider the fixed points of the function from [0, 1] to itself shown in Fig. 1. The points A and C are qualitatively similar, and qualitatively different from B. In the one-dimensional setting one can easily see that, if the function is differentiable and its graph is not tangent to the diagonal at any of its fixed points, then the number of fixed points of the first type must be one greater than the number of fixed points of the second type. In particular, the number of fixed points must be odd. These properties extend to smooth functions $f : C \to C$, where C is an n-dimensional convex set, that intersect the diagonal in the 'expected' manner: the Jacobian of $\mathrm{Id}_C - f$ is nonsingular. Debreu (1970) used Sard's theorem (for example, Milnor 1965) to show that for an exchange economy with fixed preferences, the excess demand function generated by a 'generic' endowment vector has well-behaved equilibria, and Dierker (1972) showed that the qualitative conclusions described above hold in this circumstance. Mas-Colell (1985) summarizes the extensive literature descended from these seminal contributions.

The *Leray–Schauder fixed point index* generalizes these aspects of the theory to correspondences, to sets of fixed points that are not singletons, and to general ANRs. Suppose X is a nonempty compact ANR, $U \subset X$ is open and $\overline{U}$ is its closure. A correspondence $F : \overline{U} \to X$ is *index admissible* if it is u.s.c. and does not have any fixed points in its boundary $\overline{U}/U$. Let is $\mathscr{I}_X$ be the set of index admissible contractible valued correspondences $F : \overline{U} \to X$ where $U \subset X$ is open. A homotopy $h : \overline{U} \times [0, 1] \to X$ is *index admissible* if each h_t is index admissible.

The next result gives an axiomatic characterization of a number $\Lambda_X(F)$. When there are finitely many fixed points the Additivity axiom allows us to think of $\Lambda_X(F)$ as the sum of their indices. When $X \subset \mathbb{R}^n$, $f : \overline{U} \to X$ is a smooth function, and x is a fixed point in the interior of X *with* $\mathrm{Id}_{\mathbb{R}^n} - Df(x)$ nonsingular, the index of x is $+1$ or -1 according to whether the determinant of $\mathrm{Id}_{\mathbb{R}^n} - Df(x)$ is positive or negative.



Fixed Point Theorems, Fig. 1

Theorem 6 *There is a unique function* $\Lambda_X : \mathscr{I}_X \to \mathrm{Z}$ *satisfying*:

(A) *(Normalization) If* $c : X \to X$ *is a constant function, then* $\Lambda_X(c) = 1$.

(B) *(Additivity) If* $F : \overline{U} \to X$ *is in* $\mathscr{I}_X$, $U_1, \ldots, U_r$ *are disjoint open subsets of* U, *and* F *has no fixed points in* $\overline{U}/(U_1 \cup \ldots \cup U_r)$, *then*

$$\Lambda_X(F) = \sum_{i=1}^{r} \Lambda_X\left(F_{\overline{U}_i}\right).$$

(C) *(Homotopy) If* $h : \overline{U} \times [0, 1] \to X$ *is an index admissible homotopy, then*

$$\Lambda_X(h_0) = \Lambda_X(h_1) \ .$$

(D) *(Continuity) For each* $F : \overline{U} \to X$ *in* $\mathscr{I}_X$ *there is a neighborhood* $W \subset \overline{U} \times X$ *of* $Gr(F)$ *such that* $\Lambda_X(F') = \Lambda_X(F)$ *for all* $F' : \overline{U} \to X$ *with* $F' \in \mathscr{I}_X$ *and*

$$Gr(F') \subset W.$$

The index is closed related to the Brouwer degree of a function between manifolds of the same dimension. These ideas evolved from the time of Brouwer's work until O'Neill (1953) achieved the axiomatic expression of the concept (for functions) given above.

F

Theorem 1 has many important consequences. To begin with note that if $F \in \mathscr{I}_X$ has no fixed points, then Additivity implies that

$$\Lambda_X(F) = \Lambda_X(F|_\varnothing) = \Lambda_X(F|_\varnothing) + \Lambda_X(F|_\varnothing) = 0.$$

Therefore F must have a fixed point whenever $\Lambda_X(F) \neq 0$. If $f : X \to X$ is a continuous function, then $\Lambda_X(f)$ is called the *Lefschetz* number of f. The famous Lefschetz (1923) fixed point theorem states that f has a fixed point if its Lefschetz number is nonzero, and provides connections to algebraic topology that give tools for computing the Lefschetz number.

We now use the following approximation result to recover the weak version of the Eilenberg–Montgomery theorem stated above, thereby showing that Theorem 6 embodies the fixed point principle. This result generalizes Kakutani's method of passing from Brouwer's theorem to his result, and it plays an important *role* in one method of proving Theorem 6.

Theorem 7 (Mas-Colell 1974; McLennan 1989b) *If X is a compact ANR, U, $V \subset X$ are open with $\overline{V} \subset U$, $F : \overline{U} \to X$ is a u . s . c. contractible valued correspondence, and $W \subset \overline{U} \times X$ is a neighbourhood of $Gr(F)$, then there is a continuous function $f : \overline{V} \to X$ with $Gr(f) \subset W$.*

Suppose that $F : X \to X$ is a u.s.c. contractible valued correspondence. Applying the last result with $U = V = X$ and W as in (14), we find that there is a continuous function $f : X \to X$ with $\Lambda_X(f) = \Lambda_X(F)$. If X is contractible, so that there is a homotopy $h : X \times [0, 1] \to X$ with $h_0 = Id_X$ and h_1 a constant function, then $j(x, t) = f(h(x, t))$ is a homotopy with $j_0 = f$ and j_1 a constant function, so Homotopy and Normalization imply that $\Lambda_X(f) = 1$. We conclude that $\Lambda_X(F) = 1$, and that F necessarily has a fixed point.

Recall that a subset C of a metric space Y is *connected* if there do not exist open sets V_1, $V_2 \subset Y$ with $V_1 \cap V_2 \neq \varnothing$ and $V_1 \cap C \neq \varnothing \neq V_2 \cap C$. A subset of Y is a *connected* component if it is the union of all connected sets containing some point y. Each connected component is connected, and the connected components partition Y.

Suppose that X is a compact contractible ANR, that $F : X \to X$ is in $\mathscr{I}_X$, and that the set of fixed points of F has finitely many connected components $C_1, \ldots, C_r$. Additivity implies that each component C_i has a well-defined index λ_i that depends on the restriction of F to an arbitrarily small neighbourhood of C_i. Suppose that it is possible to show that $\lambda_i = 1$ for each i. Since additivity implies that $\sum_i \lambda_i = \Lambda_X(F) = 1$, it follows that $r = 1$. This style of proof of uniqueness is applicable to many economic settings, but usually more elementary methods are available. At present no alternative to its application in Eraslan and McLennan (2005) is known. It is more common to use the index to prove nonuniqueness: it suffices to display a connected component whose index is different from one.

The fixed point index has two other important properties.

Theorem 8 (Multiplication) *If X and Y and compact ANRs, $U \subset X$ and $V \subset Y$ are open, $F : \overline{U} \to X$ and $G : \overline{V} \to Y$ are index admissible contractible valued correspondences, and $F \times G : \overline{U} \times \overline{V} \to X \times Y$ is the correspondence that takes (x, y) to $F(x) \times G(y)$, then*

$$\Lambda_{X\times Y}(F \times G) = \Lambda_X(F) \cdot \Lambda_Y(G).$$

Theorem 9 (Commutativity) *If X and Y are compact ANRs and f: $X \to Y$ and g: $Y \to X$ are continuous functions, then*

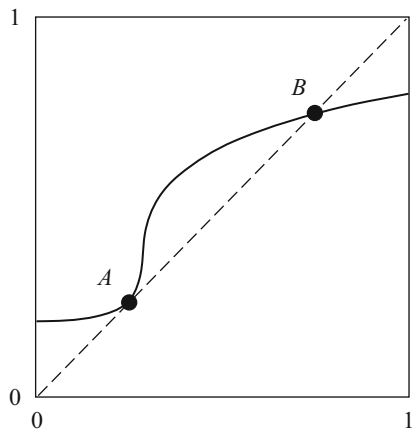
$$\Lambda_X(g\,o\,f) = \Lambda_Y(f\,o\,g).$$

There is a more general version of Commutativity for functions defined on subsets of X and Y, but its statement involves technical complications. In view of the uniqueness asserted in Theorem 6, Multiplication and Commutativity are, in principle, consequences of (A)–(D), but it is not known how to prove them in this way. In practice these properties are treated as axioms and shepherded up the ladder of generality, one rung at a time, along with everything else. In fact Commutativity (which was introduced by Browder 1948, for this purpose) plays a critical role at one stage of this process.

Essential Sets of Fixed Points

The two fixed points in Fig. 2 are qualitatively different. Arbitrarily small perturbations of the function have no fixed point near A, but this is not the case for B. In the terminology introduced by Fort (1950) A is *inessential* while B is *essential*. Let X be a compact contractible ANR, let $F : X \to X$ be a u.s.c. contractible valued correspondence, and let C be the set of fixed points of F. Kinoshita (1952) extended Fort's ideas to correspondences, and to sets of fixed points, defining an *essential set of fixed points* of F to be a compact $C' \subset C$ such that for any neighbourhood U of C' there is a neighbourhood W of Gr(F) such that any continuous function $f: X \to X$ with $Gr(f) \subset W$ has a fixed point in U.

For any neighbourhood U of C we can find a neighbourhood W of Gr(F) that cannot have any fixed points outside of U, so C is essential. That is, without some additional condition, essentiality does not distinguish some fixed points from others. Following Kohlberg and Mertens (1986), one is led to consider minimal essential sets, which exist by virtue of the following argument. Let B_1, B_2, ... be a listing of the open balls of rational radii centred at points in some countable dense subset of X. Define a sequence K_0, K_1, K_2, ... inductively by setting $K_0 = C$ and, for $j \geq 1$, setting $K_j = K_{j-1}/B_j$ if this set is essential and otherwise setting $K_j = K_{j-1}$. We claim that $K_\infty = \cap_j K_j$ is a minimal essential set.



Fixed Point Theorems, Fig. 2

Any neighbourhood U of K_∞ contains some K_j (the accumulation points of a sequence $\{x_j\}$ with $x_j \in K_j/U$ must be outside U but also in each K_j, by compactness, hence in K_∞) and each K_j is essential, so K_∞ is essential. If there was a smaller essential set there would be some j such that $K_\infty/B_j \neq K_\infty$ was essential, but then K_{j-1}/B_j would also be essential, in which case $K_\infty \cap B_j \subset K_j \cap B_j = \varnothing$.

Kinoshita (1952) showed that minimal essential sets are connected when X is convex and F is convex valued. Otherwise one could find a minimal essential set $C_1 \cup C_2$, where C_1 and C_2 are nonempty, compact, and disjoint. Then C_1 and C_2 are inessential, so there is a perturbation of F that has no fixed points near C_1 and another such perturbation of F has no fixed points near C_2. The main idea of Kinoshita's argument is that these can be combined, by using convex combination with locally varying weights, to give a perturbation of F that has no fixed point near $C_1 \cup C_2$, thereby contradicting the assumption that $C_1 \cup C_2$ is essential.

Kinoshita's theorem is pertinent to the literature on refinements of Nash equilibrium that began with the introduction in Selten (1975) of perfect equilibrium. An important technique is to give a privileged status to those Nash equilibria that can be approximated by fixed points of certain perturbations of the given correspondence. In particular, it has important connections to the notion of strategic stability of Kohlberg and Mertens (1986).

The fixed point index also has implications for essential sets. For the sake of simplicity assume that C consists of finitely many connected components $C_1, \ldots, C_r$. (This condition holds in the application to Nash equilibrium.) Any C_i with nonzero index is essential, by Continuity. Since the sum of the indices is one, some C_i must have nonzero index, so a connected essential set exists. Harder arguments, which apply the Hopf theorem (Milnor 1965) to 'transport' fixed points of perturbations to a desired location, and to eliminate pairs of fixed points of opposite index, show that any proper subset of a C_i is inessential, and that C_i is inessential if its index is zero. Thus the minimal essential sets are precisely those C_i with nonzero index.

See Also

- ▶ Computation of General Equilibria
- ▶ Computation of General Equilibria (New Developments)
- ▶ Existence of General Equilibrium
- ▶ Game Theory
- ▶ Mathematics and Economics
- ▶ Nash Equilibrium, Refinements of
- ▶ Non-Cooperative Games (Equilibrium Existence)

Bibliography

Algower, E., and K. Georg. 1990. *Numerical continuation methods*. New York: Springer Verlag.

Arrow, K., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Border, K. 1985. *Fixed point theorems with applications to economics and game theory.* Cambridge: Cambridge University Press.

Borsuk, K. 1967. *Theory of retracts*. Warsaw: Polish Scientific Publishers.

Brouwer, L. 1910. Uber Abbildung von Mannigfaltikeiten. *Mathematische Annalen* 71: 97–115.

Browder, F. 1948. *The topological fixed point theory and its applications to functional analysis*. Ph.D. thesis, Princeton University.

Brown, R. 1971. *The Lefschetz fixed point theorem*. Glenview: Scott Foresman and Co.

Debreu, G. 1952. A social equilibrium existence th. *Proceedings of the National Academy of Science* 38: 886–893.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38: 387–392.

Dierker, E. 1972. Two remarks on the number of equilibria of an economy. *Econometrica* 40: 951–953.

Doup, T. 1988. *Simplicial algorithms on the simplotope*. Berlin: Springer-Verlag.

Dugundji, J., and A. Granas. 2003. *Fixed point theory.* New York: Springer-Verlag.

Echenique, F. 2005. A short and constructive proof of Tarski's fixed point th. *International Journal of Game Theory* 33: 215–218.

Eilenberg, S., and D. Montgomery. 1946. Fixed-point theorems for multivalued transformations. *American Journal of Mathematics* 68: 214–222.

Eraslan, H., and A. McLennan. 2005. *Uniqueness of stationary equilibrium payoffs in coalitional bargaining*. Mimeo, University of Pennsylvania.

Fan, K. 1952. Fixed point and minimax theorems in locally convex linear spaces. *Proceedings of the National Academy of Sciences* 38: 121–126.

Fort, M. 1950. Essential and nonessential fixed points. *American Journal of Mathematics* 72: 315–322.

Garcia, C., and W. Zangwill. 1981. *Pathways to solutions, fixed points, and equilibria*. Englewood Cliffs: Prentice-Hall.

Glicksberg, I. 1952. A further generalization of the Kakutani fixed point theorem with applications to Nash equilibrium. *Proceedings of the American Mathematical Society* 3: 170–174.

Herings, P. 1997. An extremely simple proof of the K–K–M–S th. *Economic Theory* 10: 361–367.

Hopenhayn, H., and E. Prescott. 1992. Stochastic monotonicity and stationary distributions for dynamic economies. *Econometrica* 60: 1387–1406.

Kakutani, S. 1941. A generalization of Brouwer's fixed point th. *Duke Mathematical Journal* 8: 416–427.

Kinoshita, S. 1952. On essential components of the set of fixed points. *Osaka Mathematical Journal* 4: 19–22.

Kinoshita, S. 1953. On some contractible continua without the fixed point property. *Fundamentae Mathematicae* 40: 96–98.

Knaster, B., C. Kuratowski, and C. Mazurkiewicz. 1929. Ein Beweis des Fixpunktsatzes fur n-dimensionale Simplexe. *Fundamenta Mathematicae* 14: 132–137.

Kohlberg, E., and J.-F. Mertens. 1986. On the strategic stability of equilibria. *Econometrica* 54: 1003–1038.

Lefschetz, S. 1923. Continuous transformations of manifolds. *Proceedings of the National Academy of Sciences* 9: 90–93.

Mas-Colell, A. 1974. A note on a theorem of F. Browder. *Mathematical Programming* 6: 229–233.

Mas-Colell, A. 1985. *The theory of general economic equilibrium: A differentiable approach*. Cambridge: Cambridge University Press.

McKenzie, L. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27: 54–71.

McLennan, A. 1989a. Consistent conditional systems in noncooperative game theory. *International Journal of Game Theory* 18: 141–174.

McLennan, A. 1989b. Fixed points of contractible valued correspondences. *International Journal of Game Theory* 18: 175–184.

McLennan, A., and R. Tourky. 2005. *From imitation games to Kakutani*. Mimeo, University of Minnesota.

Milgrom, P., and C. Shannon. 1994. Monotone comparative statics. *Econometrica* 62: 157–180.

Milnor, J. 1965. *Topology from the differentiable viewpoint*. Charlottesville: University Press of Virginia.

Milnor, J. 1978. Analytic proofs of the 'hairy ball th' and the Brouwer fixed-point th. *American Mathematical Monthly* 85: 521–524.

Nash, J. 1950. *Non-cooperative games*. Ph.D. thesis, Department of Mathematics, Princeton University.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54: 286–295.

O'Neill, B. 1953. Essential sets and fixed points. *American Journal of Mathematics* 75: 497–509.

Reny, P. 2005. *On the existence of monotone pure strategy equilibria in Bayesian games*. Mimeo, University of Chicago.
Scarf, H. 1973. *The computation of economic equilibria*. New Haven: Yale University Press.
Schauder, J. 1930. Der Fixpunktsatz in Funktionalraumen. *Studia Mathematica* 2: 171–180.
Selten, R. 1975. Re-examination of the perfectness concept for equilibrium points of extensive games. *International Journal of Game Theory* 4: 25–55.
Shapley, L. 1973a. On balanced games without side payments. In *Mathematical programming study*, ed. T. Hu and S. Robinson. New York: Academic Press.
Shapley, L. 1973b. *On balanced games without side payments: A correction*, Rand paper series report no. p-4910/1. Santa Monica: RAND Corporation.
Sperner, E. 1928. Neuer Beweis fiir die Invarianz der Dimensionszahl und des Gebietes. *Abhandlungen aus dem Mathematischen Seminar der Hamburgischen* 6: 265–272.
Tarski, A. 1955. A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics* 5: 285–309.

Fixprice Models

Joaquim Silvestre

Abstract

The general competitive theory of markets (Walras, Arrow-Debreu) presupposes that no agent has market power and that prices and wages instantaneously adjust to equilibrate price-taking supply and demand. Fixprice models follow its emphasis on the interactions across markets, but under the more realistic assumption that markets frequently operate under excess demand or supply, with prices often exceeding marginal costs because prices and wages adjust slowly, or because of market power. The original fixprice models, which adopted the short-run method with static expectations, are the precursors of neo-Keynesian dynamic macroeconomics based on market power and the stickiness of wages or prices.

Keywords

Bargaining; Comparative statics; Dual decision hypothesis; Dynamic stochastic macroeconomic models; Employment; Excess demand and supply; Fixprice models; Full employment; General equilibrium; Imperfect competition; Inflation; Market power; Microfoundations; Monopolistic competition; Oligopoly; Oligopsony; Patinkin, D.; Price control; Real business cycles; Rent control; Second best; Staggered prices; Staggered wages; Sticky prices; Sticky wages; Unemployment; Voluntariness; Wage control; Walrasian equilibrium; Walras–Samuelson tâtonnement

JEL Classifications

F3

The canonical fixprice model (Benassy 1975, 1976, 1982; Drèze 1975; Younès 1975) was born at the interface of two extensions of general equilibrium theory: the study of out-equilibrium price dynamics, and the incorporation of price-setting behaviour by firms. Fixprice analysis aimed at providing microfoundations for macroeconomic theory and policy. Accordingly, it first generated static macroeconomic models of the interaction between the labour and the output markets at given prices and wages with or without explicit market power. Later, it exerted a diffuse influence on the more recent dynamic macroeconomic models with market power and/or wage or price stickiness.

Many wages and prices appear to change infrequently and fail to respond quickly to shocks. Casual observation suggests that the wages of many workers are fixed in nominal terms for at least several months, and do not drop quickly in response to adverse shocks in demand. This observation is well supported by quantitative studies (Taylor 1999) as well as by the in-depth interviews of Bewley (1999). For instance, Cecchetti (1984) found that, even when the rate of inflation was high, union wages were fixed at

nominal levels for an average of one year. Later researchers, such as Card and Hyslop (1997), have obtained similar results for non-union workers. Bewley (1999) finds that wage rigidity is stricter in long-term, full-time jobs (the 'primary sector') than in short-term, part-time ones, and emphasizes downwards rigidity over upwards rigidity, an asymmetry that is qsted by Taylor (1999).

Price rigidity has also been subject to extensive inquiry (Andersen 1994; Taylor 1999). Even though one may presume that prices are less rigid than wages, and many commodities are indeed sold at continuously changing prices, several studies have found that, on the average, prices may stay fixed for relatively long periods (Carlton 1986, 1989; Cecchetti 1986; Blinder et al. 1998). In Taylor's (1999, p. 1020) words, '. . . the studies suggest that *price changes and wage changes have about the same average frequency – about one year*' (emphasis in original.) Later work has found shorter, but still ample, average periods (Baharad and Eden 2004).

In addition, prices seem to systematically exceed marginal costs in many industrial markets. These observations challenge the relevance of models where wages and prices are assumed to adjust instantaneously to their Walrasian equilibrium values.

Theoretical Roots of the Canonical Fixprice Model

Out-of-Equilibrium Price Dynamics

The Walrasian approach postulates that prices adjust very rapidly in response to excess demand or supply, so that no transactions occur before equilibrium is reached. A rigorous formulation of this idea is the Walras–Samuelson tâtonnement process. Consider an exchange economy with two commodities and two traders, Trader i being initially endowed with ω_{ij} units of commodity $j(i, j = 1, 2)$. Let the aggregate Walrasian excess demand functions be $z^j(p_1, p_2|\omega), j = 1, 2$. (Here the vector $\omega = (\omega_{ij})$ is fixed). As long as $z^1(p_1, p_2|\omega) \neq 0$ or $z^2(p_1, p_2|\omega) \neq 0$, no transactions occur and prices adjust according to the differential equation:

$$\frac{dp_j}{dt} = z^j(p_1(t)), p_2(t)|\omega), j = 1, 2.$$

Walrasian excess demands provide the 'market signals' for the adjustment of prices in the Walras–Samuelson tâtonnement. Of course, the Walrasian excess demand functions express plans made under the conjecture that any quantities can be bought and sold at the going prices. If transactions did occur at non-Walrasian prices, then such a conjecture would be falsified, since some agents would be unable to realize their plans (see Arrow 1959). This led Patinkin to postulate that disequilibrium transactions in a market create spillover effects on others, so that, for example, 'the pressure of excess demand in the one market affects the price movements in all other markets' (1956, p. 157). Patinkin's formulation was imprecise (Negishi 1965; Clower 1965), but his search for the 'relevant market signals' motivates Clower's (1965) 'dual decision hypothesis'. This idea, generalized by Barro and Grossman (1971, 1976), is central to Benassy's fixprice model.

It was discovered in the late 1950s that the Walras–Samuelson tâtonnement process fails to converge unless some restrictive assumptions are imposed, motivating the non-tâtonnement adjustment process. Here two simultaneous movements occur: the distribution of the endowments changes according to some rule for trading at non-Walrasian prices, and prices adjust in response to Walrasian excess demands at the current endowments, for example, for some rule g_{ij},

$$\frac{d\omega_{ij}}{dt} = g_{ij}(p_1(t)), p_2(t)|\omega(\mathrm{t})\big),$$

$$\frac{dp_j}{dt} = z^j(p_1(t)), p_2(t)|\omega(t)\big), \mathrm{i,j} = 1, 2.$$

This process is hard to interpret except possibly as depicting the sequential exchange of durable goods, and, as just argued, the appeal to Walrasian excess demand in the price-adjustment equation is unjustified. But some conditions on the functions g_{ij} (Hahn and Negishi 1962; Uzawa 1962) originally meant to guarantee the convergence of the non-tâtonnement process inspired the fixprice model of Younès (1975).

General Equilibrium with Market Power

The monopolistic general equilibrium analysis pioneered by Negishi (1960) led to the construction of simple models where firms or workers had price- or wage-setting capacity (Benassy 1976, 1977, 1982, 1991; Hart 1982; Silvestre 1990, 1993). This work revealed intimate connections between market power and the fixity or stickiness of prices and wages. First, oligopoly displays formal parallelisms to excess supply, and oligopsony to excess demand (Madden and Silvestre 1991, 1992; Silvestre 1986). Second, an imbalance between supply and demand gives temporary market power to agents on the short side who then face non-horizontal demand or supply curves for large enough quantities (Arrow 1959; Negishi 1974, 1979; Hahn 1978; John 1985). Third, a firm with market power experiencing frequent demand or productivity shocks may optimize by changing prices at discrete intervals even if the costs of changing prices are small (Sheshinski and Weiss 1977; Akerlof and Yellen 1985; Mankiw 1985; Parkin 1986; Caplin and Spulber 1987; Blanchard and Kiyotaki 1987), and the resulting stickiness is magnified by strategic complementarities among the pricing decisions of firms (Fishman and Simhon 2005).

Fixprice Allocations

Trading at Non-Walrasian Prices

Fixprice analysis postulates a common medium of exchange (money) in each market. Thus, there are $n + 1$ goods (from 0 to n) in the case of n markets, the zero good being money. The analysis addresses two qsts. First, given a price vector p (normalized with respect to money), what allocations are compatible with it? Second, given a p and an allocation compatible with it, which is the type of disequilibrium in each market? The answers are derived from three basic principles: (*a*) voluntary trading; (*b*) absence of market frictions, and (*c*) effective demand. The last requires the explicit recognition of the interaction among markets. The first two impose conditions on the trades carried out in a market, namely, that, at the going price, (*a*) no trader may gain by trading less; (*b*) no pair formed by a buyer and a seller may gain by trading more.

The fixprice model provides a general framework (which includes Walrasian markets as a limit) for price-guided allocation mechanisms. It has several applications: (*a*) *short-run analysis*, which assumes that it takes time for prices and quantities to adjust; (*b*) *market power* (imperfect or monopolistic competition); (*c*) *price* (*wage or rent*) *controls*; this in particular motivates Drèze's formulation; and (*d*) *price* (*or wage*) *negotiation* (representatives of buyers and sellers negotiate prices that are taken as given by individual traders: see Silvestre 1988). Fixprice analysis can be viewed as abstracting from specific features and focusing instead on basic market principles common to alternative specifications.

The definitions of fixprice equilibrium due to Bénassy, Drèze and Younès vary in form and motivation, but turn out to be equivalent under some assumptions (Silvestre 1982, 1983). Rather than reproducing them in all generality, we exemplify the common concepts in two simple but important cases.

Differentiable Exchange Economies

There are $n + 1$ goods, indexed 0, 1,. . ., n (i.e. n markets). There are m traders: trader i is endowed with an $(n + 1)$ dimensional vector ω_i of initial endowments and a differentiable utility function $u_i(x_{i0}, x_{i1},\ldots, x_{in})$. A net trade allocation is an m-tuple of n-dimensional net trade vectors $(z_i) = (z_{i1},\ldots, z_{in})$), one for each trader, satisfying: $\Sigma_{izi} = 0$. It is understood that, for $j = 1,\ldots, n$, if $z_{ij} > 0$ (or < 0) then trader i is buying (or selling) in market j. The (normalized) price vector $p = (p_1,\ldots,p_n)$ is given. The vector $\hat{x}(p, z_i) \equiv (\omega_{i0} - p \bullet z_i, \omega_{i1} + z_{i1},\ldots, \omega_{in} + z_{in})$ is then the consumption vector associated with (p, z_i). Define i's *marginal utility of trading in market j at the going price* as: $\mu_{ij}(p, z_i) \equiv \partial u_i/\partial x_{ij} - p_j \partial u_i/\partial x_{i0}$, with derivatives evaluated at $\hat{x}_i(p, z_i)$.

Definition 1 A net trade allocation (z_i^*) is a *fixprice equilibrium for p* if, writing $\mu_{ij}^* \equiv \mu_{ij}(p, z_i^*)$:

(a) *Voluntariness*: For $i = 1, \ldots, m$, $z_{ij}^* \cdot \mu_{ij}^* \geq 0$;
(b) *Absence of market frictions*: For $j = 1, \ldots, n$ and for any pair of consumers $i, h, \mu_{ij}^* \cdot \mu_{hj}^* > 0$.

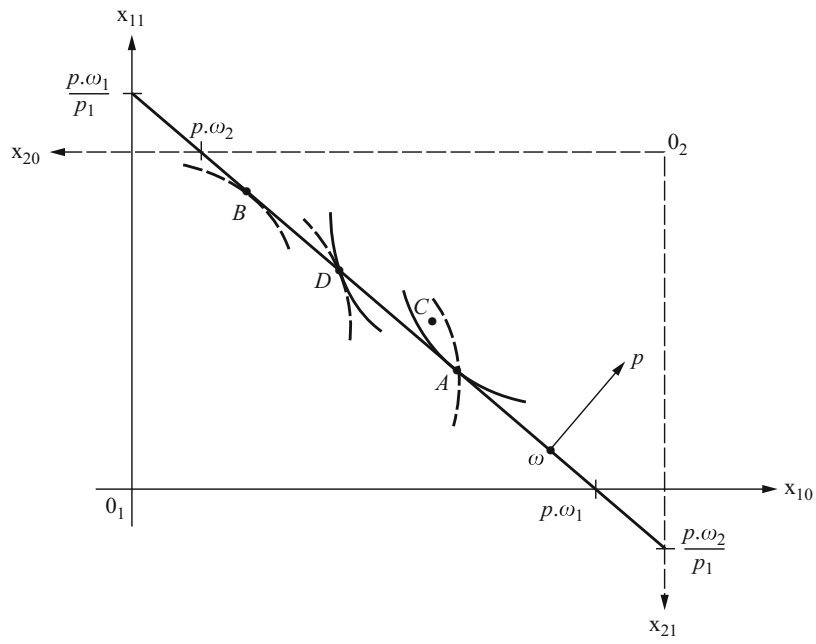
Figure 1 illustrates the case of $n = 1$ and $m = 2$ in an Edgeworth box: point A represents the (unique) fixprice equilibrium at the price vector p: there Trader 1 is a buyer ($z_{11} > 0$). The straight line through points ω and A depicts the budget constraints. Allocations in the segment $[\omega, A)$ violate condition (b). Those in the segment $[A, B)$ (in particular the Pareto efficient point D) violate condition (a) for Trader 1.

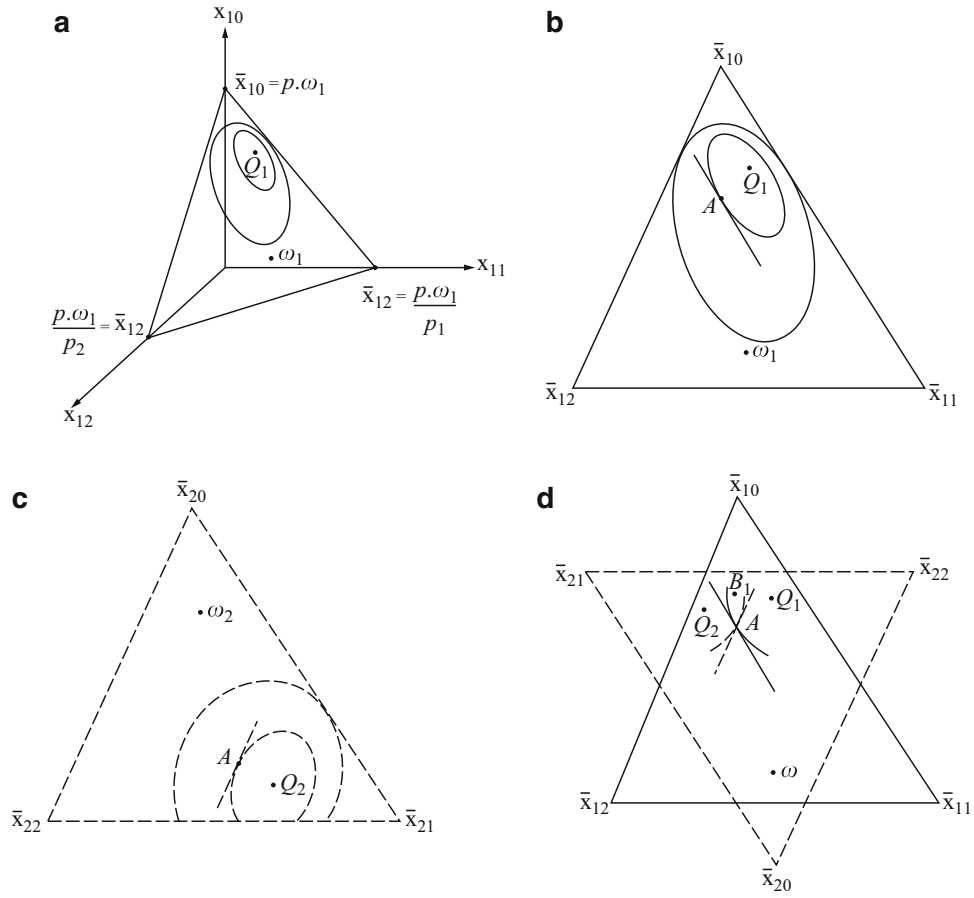
The more complex case of $n = m = 2$ is illustrated in Fig. 2a–d. Figure 2a depicts Trader 1's budget set in Re^3_+. Rather than drawing a three-dimensional Edgeworth box, we graph first separately (Fig. 2b, c) and then together (Fig. 2d) the two-dimensional budget triangles of the traders. Figure 2a, b also depicts the intersections of some indifference surfaces of Trader 1 with her budget set, Q_1 being her most preferred point in the budget set. At point A she is selling in both markets (that is, $z_{11} < 0$ and $z_{12} < 0$: she gets money in exchange), with $\mu_{12} < 0$ (she would like to sell more in market 2) and $\mu_{11} < 0$. Figure 2c corresponds to Trader 2: at point A she is buying in both markets (i.e., $z_{21} > 0$ and $z_{22} > 0$), with $\mu_{21} > 0$ and $\mu_{22} = 0$. Figure 2d superimposes the two graphs (with the axes corresponding to Trader 2 reversed, and with the initial endowment points coinciding at ω). Points A in Fig. 2b, c have been chosen so that they also coincide in Fig. 2d, i.e., $z_{2j} + z_{2j} = 0$, $j = 1, 2$. These trades constitute a fixprice equilibrium.

The Three-Good Model

The model originated in Barro and Grossman (1971, 1976) and was further elaborated by Benassy (1977, 1982, 1986) and Malinvaud (1977) among others. Let there be three goods: money, denoted by M, initially available in M_0 units; labour, denoted by L, initially available in L_0 units, and output, denoted by Y, which is produced by labour according to the production function $Y = f(L)$. There are two markets, the labour market, with (nominal) wage w, and the output market, with price p. There is one firm and one consumer, with preferences represented by the homogeneous utility function $U(Y, M)$, who

Fixprice Models, Fig. 1





Fixprice Models, Fig. 2

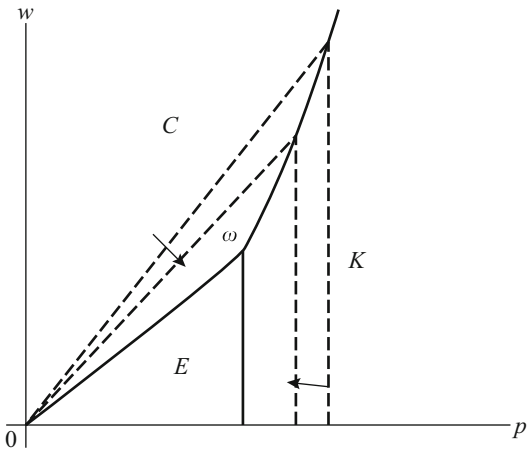
owns M_0 and L_0, and receives all profits. The labour supply is fixed at L_0.

Define the *marginal rate of substitution* as $V'(Y) \equiv \frac{\partial U / \partial Y}{\partial U / \partial M}$, with derivatives evaluated at (Y, M_0). Define the marginal cost curve as $C'_w(Y) \equiv w\left(f^{-1}\right)'(Y)$, and the *full employment output* as $Y_0 \equiv f(L_0)$.

Definition 2 The level of output Y is a fix price equilibrium output for the price–wage pair (p, w) if $Y = \min\{(V') - 1(p), (C'w) - 1(p), Y_0\}$.

This equality embodies in a compact way four conditions. First, $Y \leq Y_0$, that is, output cannot exceed the full employment level. Second, $Y \leq (V') - 1(p)$, or alternatively $p \leq V'(Y)$: the consumer cannot gain by buying less output at the going price (it is a condition of voluntary trading for the consumer). Third, $Y \leq \left(C'_w\right)^{-1}(p)$, or $p \geq C'_w(Y)$: the price cannot be lower than the marginal cost, or, in other words, profits cannot increase by selling less at the going price (it is a condition of 'voluntary trading' for the firm). Finally, at least one of these weak inequalities must be an equality: this is the condition of frictionless markets.

Figure 3 partitions the (p, w) plane according to which one of the three possible equalities determines output (solid lines). In region E (full employment), $Y = Y0$. In region K (Keynesian unemployment), $p = V'(Y)$ and in region C (classical unemployment of full capacity), $p = C'_\omega(Y)$. In the boundaries between regions the two relevant equalities hold. At the Walrasian point W all three equalities hold. There is full employment in region



Fixprice Models, Fig. 3

E and unemployment outside it. The dashed lines are isoemployment loci, with the arrows indicating the directions of increasing employment.

The labour market is in excess supply (or excess demand) in the interior of regions K and C (or region E), and the output market is in excess supply (or demand) in the interior of region K (or regions C and E). At the Walrasian point W both markets are balanced. In the Keynesian region the condition $p = V'(Y)$ for determination of output can be rewritten in terms of the consumption function as in the textbook Keynesian multiplier model. The homogeneity of U implies that demand for output, as a function of p and wealth I, can be written as $h(p)I$, where the function $h(p)$ satisfies: (a) $ph(p) < 1$ and (b) the marginal equality $(\partial U/\partial Y)/(\partial U/\partial Y) = p$ whenever the consumption vector is a multiple of $(h(p), 1 - ph(p))$. By setting $I = M_0 + pY$ we obtain the *effective demand for output* $C(Y) = h(p)(M_0 + pY)$: this is the traditional consumption function, with marginal propensity to consume equal to $ph(p) < 1$. The satisfaction of effective demand requires $Y = C(Y)$, that is, $Y/M_0 = h(p)/[1 - ph(p)]$, which by the above marginal equality implies that $p = V'(Y)$.

The distinction between the two types of excess supply of labour has important implications for policy and for comparative statics. Output is determined in region C by the condition 'price = marginal cost'. Hence, lowering wages (nominal or real) will increase employment, whereas an increase in demand will have no effect on employment. But in region K a decrease in the nominal wage has no effect on employment: only lowering the price or otherwise stimulating demand will work. This analysis also offers insights on the effects of different kinds of shocks (Malinvaud 1977): a business cycle driven by demand shocks will fluctuate between Keynesian unemployment and full employment, whereas productivity shocks will yield fluctuations between the Keynesian and the classical types of unemployment.

Welfare Analysis

The budget equality and the market institution impose constraints on trades. Thus, the resulting allocation may very well be Pareto dominated by other allocations that do not satisfy these constraints. The study of such inefficiencies is important for the normative analysis of the situations covered by fixprice theory (short-run market disequilibria, price controls, monopolistic market power). On the other hand, to the extent that these constraints cannot be circumvented, they are for policy purposes as effective as physical and resource constraints, motivating the study of efficiency subject to these additional constraints ('second best'.)

Inefficiency Relative to the Set of Physically Attainable Allocations

Consider Fig. 1. Note that the allocation given by A is not Pareto efficient: both traders would be better off at C, but C cannot be reached without violating some budget constraint.

A similar phenomenon may occur if there are two traders in one side of the market. Modify the example of Fig. 1 by duplicating Trader 2: that is, Traders 1 and 2 are unchanged, but now there is a Trader 3 with the same preferences and endowments as Trader 2. Let $z_{21} = (-1/4)z_{11}$, and $z_{31} = (-3/4)z_{11}$. Then there are mutually beneficial reallocations between Traders 2 and 3, but they violate the budget constraint.

One can say that this type of inefficiency is caused by 'wrong prices'. Note, however, that trade at non-Walrasian prices does not per se imply inefficiency. Point D in Fig. 1, for instance, is Pareto efficient, and all budget constraints are satisfied there. (A general treatment of allocations of this type is given by Balasko 1979, and Keiding 1981). But there is forced trading at point D. It is the combination of non-Walrasian prices and the voluntariness condition that implies inefficiency (Silvestre 1985).

Inefficiencies Relative to Allocations Satisfying the Budget Constraint

When there is only one market (see Fig. 1), the absence of frictions guarantees that no allocation that satisfies the budget constraint is Pareto superior to a fixprice equilibrium; that is, a fixprice equilibrium is efficient relative to allocations that satisfy the budget constraints. This ceases to be true with several markets: for instance, point B in Fig. 2d Pareto dominates point A and satisfies all budget constraints. (Note that point B violates voluntariness.) Such inefficiencies have been studied in Benassy (1975, 1977, 1982) and Younès (1975). A particularly striking case occurs in Keynesian allocations of the three-good model: the markets for labour and output are in excess supply, and a direct barter of labour against output would benefit both the firm and the worker, and improve welfare. This phenomenon was viewed by Clower (1965) as a failure of coordination among markets.

Undominated Price–Wage Pairs

Suppose that, in the three-good model, wages and prices are determined by negotiation between representatives of labour and business, and then taken as given by individual firms and workers, so that a fixprice allocation results. If bargaining is efficient, any movement away from the negotiated price–wage pair (p, w) will make somebody worse off, in which case we say that (p, w) is *undominated*. Do there exist undominated price–wage pairs besides the Walrasian pair? Note that this question is different from the ones addressed in the previous paragraphs: there, we compared allocations at a given (p, w), whereas now we compare price–wage pairs.

The answer depends on the rationing of unemployment, that is, on whether unemployment falls uniformly on workers, or, on the contrary, some workers are dismissed whereas others experience no rationing (Silvestre 1988, 1989). In the first case, the answer is affirmative under some assumptions, in which case the set of undominated price–wage pairs is a segment of the Keynesian–classical boundary of Fig. 3, implying that the output market is balanced. Non-uniform rationing of unemployment typically expands this set to a band of the Keynesian region adjoining the Keynesian–classical boundary.

The analysis is extended to a dynamic model by Jacobsen and Schultz (1990, 1991), who characterize the conditions for unemployment at undominated wages under the assumptions that unemployment is uniformly rationed, and that the output market is always balanced.

Wage and Price Rigidities in Dynamic Macroeconomic Models

Dynamic stochastic macroeconomic models were first developed under the Walrasian assumptions of price taking and market clearing in models labelled 'real business cycle' (Kydland and Prescott 1982; King and Plosser 1984) aimed at mimicking business cycle regularities. Later developments have improved the fit, in particular for the persistence of real effects of monetary shocks, by introducing, singly or in combination, market power, or price or wage rigidity.

Here we focus on rigidities. (See Silvestre 1995, for an early account of market-power, dynamic macroeconomic models; Svensson 1986, combines market power with sticky prices in a dynamic model). A first departure from Walrasian market clearing is the assumption that nominal wages are predetermined in the short run, before technological or monetary shocks are experienced. For instance, they may be preset at the expected Walrasian level (Gray 1976), so that

expected demand equals expected supply, whereas actual discrepancies between supply and demand are resolved in favour of demand: workers supply the amount of labour demanded by firms at the predetermined wage. (This simplifying assumption conflicts with the voluntariness condition of the canonical fixprice model, as described above. Benassy 1995b, 2002, modifies the dynamic model of preset wages by postulating that unions maximize a utility function subject to the voluntariness condition.) This form of rigidity or stickiness yields predictions quite different from the Walrasian model: it grants monetary shocks the ability to generate large effects on employment and output, allowing for contemporary money shocks to generate countercyclical behaviour of the real wage, as well as cyclical behaviour of prices (Benassy 1995a). The determinants of the accompanying welfare costs are studied in Cho et al. (1997). A shortcoming of this approach is that monetary shocks show relatively little persistence, limited by the length of the period in which wages are fixed (Taylor 1999).

This limitation, together with the observation that wage contracts are not synchronized across firms, led to the models of staggered wages or prices. In their simplest form (Taylor 1979, 1980), wages are fixed for a given number N of dates, but in each date $1/N$ firms renew their contracts, so that at any moment the average wage is defined by the current contract plus the ones set in the last $N - 1$ dates. More complex versions allow for various contract lengths. An influential formulation is that of Calvo (1983), who postulates that the contract length is stochastic: a given contract remains unchanged at each date with probability π, and terminated and reset with probability $1 - \pi$. This approach typically yields propagation mechanisms and persistence characteristics capable of matching stylized facts of economic fluctuations (Benassy 2002, 2003; Yun 1996). Christiano et al. (2005) show that wage staggering performs better than price staggering in generating the observed type of persistence, confirming Andersen's (1998) analysis in a model based on staggered prices or wages with fixed contract length.

See Also

- Arrow–Debreu Model of General Equilibrium
- Dynamic Models with Non-clearing Markets
- General Equilibrium
- Keynesianism
- Microfoundations
- New Keynesian Macroeconomics
- Second Best
- Sticky Wages and Staggered Wage Setting
- Temporary Equilibrium
- Underemployment Equilibria
- Unemployment

Bibliography

Akerlof, G.A., and J.L. Yellen. 1985. A near-rational model of the business cycle, with wage and price inertia. *Quarterly Journal of Economics* 100: 823–838.

Andersen, T.M. 1994. *Price rigidity: Causes and macroeconomic implications*. Oxford: Clarendon Press.

Andersen, T.M. 1998. Persistence in sticky price models. *European Economic Review* 42: 593–603.

Arrow, K.J. 1959. Towards a theory of price adjustment. In *The allocation of economic resources*, ed. M. Abramovitz. Stanford: Stanford University Press.

Baharad, E., and B. Eden. 2004. Price rigidity and price dispersion: Evidence from micro data. *Review of Economic Dynamics* 7: 613–641.

Balasko, Y. 1979. Budget constrained Pareto-efficient allocations. *Journal of Economic Theory* 21: 359–379.

Barro, R.J., and H.I. Grossman. 1971. A general equilibrium model of income and employment. *American Economic Review* 61: 82–93.

Barro, R.J., and H.I. Grossman. 1976. *Money, employment and inflation*. Cambridge: Cambridge University Press.

Benassy, J.P. 1975. Neo-Keynesian disequilibrium theory in a monetary economy. *Review of Economic Studies* 42: 503–523.

Benassy, J.P. 1976. The disequilibrium approach to monopolistic price setting and general monopolistic equilibrium. *Review of Economic Studies* 43: 69–81.

Benassy, J.P. 1977. A neo-Keynesian model of price and quantity determination in disequilibrium. In *Equilibrium and disequilibrium in economic theory*, ed. G. Schwödiauer. Boston: Reidel.

Benassy, J.P. 1982. *The economics of market disequilibrium*. New York: Academic Press.

Benassy, J.P. 1986. *Macroeconomics: An introduction to the non-Walrasian approach*. New York: Academic Press.

Benassy, J.P. 1991. Microeconomic foundations and properties of a macroeconomic model with imperfect competition. In *Issues in contemporary economics: Markets and welfare*, ed. K.J. Arrow, vol. 1. London: Macmillan.

Benassy, J.P. 1995a. Money and wage contracts in an optimizing model of the business cycle. *Journal of Monetary Economics* 35: 303–315.

Benassy, J.P. 1995b. Nominal rigidities in wage-setting by rational trade unions. *Economic Journal* 105: 635–643.

Benassy, J.P. 2002. *The macroeconomics of imperfect competition and nonclearing markets: A dynamic general equilibrium approach*. Cambridge, MA: MIT Press.

Benassy, J.P. 2003. Staggered contracts and persistence: Microeconomic foundations and macroeconomic dynamics. *Recherches Économiques de Louvain/Louvain Economic Review* 69: 125–144.

Bewley, T.F. 1999. *Why wages don't fall during a recession*. Cambridge, MA: Harvard University Press.

Blanchard, O.J., and N. Kiyotaki. 1987. Monopolistic competition and the effects of aggregate demand. *American Economic Review* 77: 647–666.

Blinder, A.S., E.R. Canetti, D.E. Lebow, and J. Rudd. 1998. *Asking about prices: A new approach to understanding price stickiness*. New York: Russell Sage Foundation.

Calvo, G.A. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12: 383–398.

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

Card, D., and D. Hyslop. 1997. Does inflation 'grease the wheels of the labor market'? In *Reducing inflation*, ed. C. Romer and D. Romer. Chicago: University of Chicago Press.

Carlton, D.W. 1986. The rigidity of prices. *American Economic Review* 76: 637–658.

Carlton, D.W. 1989. The theory and the facts of how markets clear: Is industrial organization valuable for understanding macroeconomics? In *Handbook of industrial organization*, ed. R. Schmalensee and R.D. Willig, vol. 1. Amsterdam: North-Holland.

Cecchetti, S.G. 1984. Indexation and incomes policy: A study of wage adjustment in unionized manufacturing. *Journal of Labor economics* 5: 391–412.

Cecchetti, S.G. 1986. The frequency of price adjustment: A study of newsstand prices of magazines. *Journal of Econometrics* 31: 255–274.

Cho, J.O., T. Cooley, and L. Phaneuf. 1997. The welfare cost of nominal wage contracting. *Review of Economic Studies* 64: 465–484.

Christiano, L.J., M. Eichenbaum, and C.L. Evans. 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113: 1–45.

Clower, R.W. 1965. The Keynesian counter-revolution: A theoretical appraisal. In *The theory of interest rates*, ed. F.H. Hahn and F.P.R. Brechling. London: Macmillan.

Drèze, J. 1975. Existence of an exchange equilibrium under price rigidities. *International Economic Review* 16: 301–320.

Fishman, A., and A. Simhon. 2005. Can small menu costs explain stickyprices? *Economics Letters* 87: 227–230.

Gray, J.A. 1976. Wage indexation: A macroeconomic approach. *Journal of Monetary Economics* 2: 221–235.

Hahn, F.H. 1978. On non-Walrasian equilibria. *Review of Economic Studies* 45: 1–17.

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

Hart, O. 1982. A model of general equilibrium with Keynesian features. *Quarterly Journal of Economics* 97: 109–138.

Jacobsen, H.J., and C. Schultz. 1990. A general equilibrium, macro model with wage bargaining. *Scandinavian Journal of Economics* 92: 379–398.

Jacobsen, H.J., and C. Schultz. 1991. Undominated wage rates in a unionized, overlapping generations economy. *European Economic Review* 35: 1255–1275.

John, R. 1985. A remark on conjectural equilibria. *Scandinavian Journal of Economics* 87: 137–141.

Keiding, H. 1981. Existence of budget constrained Pareto-efficient allocations. *Journal of Economic Theory* 24: 393–397.

King, R.G., and C.I. Plosser. 1984. Money, credit and prices in a real business cycle. *American Economic Review* 74: 363–380.

Kydland, F.E., and E. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.

Madden, P., and J. Silvestre. 1991. Imperfect competition and fixprice equilibria when goods are gross substitutes. *Scandinavian Journal of Economics* 93: 479–494.

Madden, P., and J. Silvestre. 1992. Imperfect competition and fixprice equilibria under consumer aggregation and net substitutes. *Scandinavian Journal of Economics* 94: 103–111.

Malinvaud, E. 1977. *The theory of unemployment reconsidered*. Oxford: Basil Blackwell.

Mankiw, N.G. 1985. Small menu costs and large business cycles: A macroeconomic model of monopoly. *Quarterly Journal of Economics* 100: 529–539.

Negishi, T. 1960. Monopolistic competition and general equilibrium. *Review of Economic Studies* 28: 196–201.

Negishi, T. 1965. Market clearing processes in a monetary economy. In *The theory of interest rates*, ed. F.H. Hahn and F.P.R. Brechling. London: Macmillan.

Negishi, T. 1974. Involuntary unemployment and market imperfection. *Economic Studies Quarterly* 25: 32–46.

Negishi, T. 1979. *Microeconomic foundations of Keynesian macroeconomics*. Amsterdam: North-Holland.

Patinkin, D. 1956. *Money, interest and prices*. Evanston: Row Peterson.

Parkin, M. 1986. The output–inflation tradeoff when prices are costly to change. *Journal of Political Economy* 94: 200–224.

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

Silvestre, J. 1982. Fixprice analysis in exchange economies. *Journal of Economic Theory* 26: 28–58.

Silvestre, J. 1983. Fixprice analysis in productive economies. *Journal of Economic Theory* 30: 401–409.

Silvestre, J. 1985. Voluntary and efficient allocations are Walrasian. *Econometrica* 53: 807–816.
Silvestre, J. 1986. The elements of fixprice microeconomics. In *Microeconomic theory*, ed. L. Samuelson. Boston: Kluwer Nijhoff.
Silvestre, J. 1988. Undominated prices in the three good model. *European Economic Review* 32: 161–178.
Silvestre, J. 1989. Who benefits from unemployment? In *The economics of imperfect competition and employment: Joan Robinson and beyond*, ed. G.R. Feiwel. London: Macmillan.
Silvestre, J. 1990. There may be unemployment when the labor market is competitive and the output market is not. *Economic Journal* 100: 899–913.
Silvestre, J. 1993. The market-power approach to macroeconomic policy. *Journal of Economic Literature* 31: 105–141.
Silvestre, J. 1995. Market power and macroeconomic policy: New developments. *Annales d'Économie et de Statistique* 37–38: 319–356.
Svensson, L. 1986. Sticky goods prices, flexible asset prices monopolistic competition and monetary policy. *Review of Economic Studies* 53: 385–405.
Taylor, J.B. 1979. Staggered wage setting in a macro model. *American Economic Review* 69: 108–113.
Taylor, J.B. 1980. Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88: 1–23.
Taylor, J.B. 1999. Staggered price and wage setting in macroeconomics. In *Handbook of macroeconomics*, ed. J.B. Taylor and M. Woodford, vol. 1. Amsterdam: North-Holland.
Uzawa, H. 1962. On the stability of Edgeworth's barter process. *International Economic Review* 3: 218–232.
Younès, Y. 1975. On the role of money in the process of exchange and the existence of a non-Walrasian equilibrium. *Review of Economic Studies* 42: 489–501.
Yun, T. 1996. Nominal price rigidity, money supply endogeneity, and business cycles. *Journal of Monetary Economics* 37: 345–370.

Fleming, John Marcus (1911–1976)

S. C. Tsiang

Keywords
Fleming, J. M.; Haberler, G. H.; Welfare economics

JEL Classifications
B31

Fleming was born on 13 March 1911 at Bathgate, Scotland, and died on 3 February 1976. He was educated at Edinburgh University, where he received the degrees of MA (Honours) in history in 1932 and MA (First Class Honours) in political economy in 1934. He was a graduate research fellow at the Institut Universitaire des Hautes Etudes Internationales in 1934–5, and a graduate student at the London School of Economics in 1935.

At the end of 1935, he joined the Secretariat of the League of Nations, Economic Intelligence Section, as a research economist, and assisted Gottfried Haberler in the latter's *Prosperity and Depression* (first published by the League in 1937). During the Second World War, he served with the UK Ministry of Economic Warfare from 1939 until 1942, and then joined the Economic Section of the Cabinet Office under Lord Robbins, rising eventually to the position of Deputy Director of the section. He was also a member of the UK Delegation to the San Francisco Conference in 1945; a member of the Preparatory Commission of the United Nations in 1946; a member of the International Trade Conference Preparatory Commission, 1947; and UK Representative to the Economics and Employment Commission, United Nations in 1950. From 1951 to 1954 Fleming was Visiting Professor at Columbia University, New York. He joined the International Monetary Fund in 1954 as a division chief, and in 1964 became the Deputy Director of the Research Department.

His academic contributions are mostly in the fields of welfare theory and trade and exchange policies. The most notable of his contributions was the seminal article 'On Making the Best of Balance of Payments Restrictions on Imports' (1951), which, in James Meade's words, was 'the begetter of the analysis of the second best' (Meade 1978), which rapidly became a fashionable new topic in welfare theory during the 1950s.

Selected Works

1944. (With J.E. Meade.) Price and output policy of state enterprise: A symposium. *Economic Journal* 54: 321–8, 337–9.

1951. On making the best of balance of payments restrictions on imports. *Economic Journal* 61: 48–71.
1952. A cardinal concept of welfare. *Quarterly Journal of Economics* 6: 366–84.
1961. International liquidity: Ends and means. *IMF Staff Papers* 8: 439–63.
1962. Domestic financial policies under fixed and under floating exchange rates. *IMF Staff Papers* 9: 369–79.
1964. (With R.A. Mundell.) Official intervention in the forward exchange market. *IMF Staff Papers* 11: 1–17.
1971a. *Essays in international economics.* Cambridge, MA: Harvard University Press.
1971b. The SDR: Some problems and possibilities. *IMF Staff Papers* 18: 25–47.
1977. International aspects of inflation. In *Inflation theory and anti-Inflation policy,* ed. E. Lundberg. London: Macmillan.

Bibliography

Haberler, G. 1937. *Prosperity and depression*. Geneva: League of Nations.
Meade, J.E. 1978. Commemorations. In *Essays on economic policy*, ed. J. Marcus Fleming (posthumous). New York: Columbia University Press.

Flexible Exchange Rates

R. Driskill

Flexible exchange rates are market determined prices of foreign exchange which move in response to supply and demand and are not pegged within narrow bands by official purchases. Flexible systems where there are no official purchases are usually called pure floating regimes, and systems with some official purchases are called managed floating regimes. The counterpart to flexible exchange rates are fixed exchange rates, where official central bank purchases or sales of foreign currencies maintain the exchange rate within narrow bands.

Until 1973, historical experience with flexible exchange rates was limited, and seemed incapable of telling much about general principles of flexible rate systems. The best known experiences involved short periods of floating brought about by the collapse of fixed rate systems. During the 1920s and 1930s, sporadic floating occurred when exchange-rate pegging failed in response to severe real and financial upheavals. These experiences, characterized by apparently destabilizing speculation and highly variable exchange rates, left many central bankers, businessmen, and a few economists, wary about the efficacy of floating rates during more tranquil periods.

In contrast to central bankers, academic economists were more sanguine about the operating characteristics of floating exchange rates. As is often the case in monetary economics Milton Friedman early and persuasively made the case for flexible exchange rates with his article, 'The Case for Flexible Exchange Rates'. He pointed out the macroeconomic benefits from flexible exchange rates, especially monetary independence, and argued that speculators, the purported villains of the interwar floating experiences, would actually help ensure the smooth working of floating rates. By the time of the breakdown of the Bretton Woods system of fixed rates, most academic economists advocated flexible rates and felt that such rates would not be very volatile.

In fact, experience since 1973 shows us that our theoretical musings about how flexible rates would work were quite wrong: exchange rates have been extremely volatile. The post-1973 events, though, have stimulated international economists to develop new theories about exchange-rate determination. These theories emphasize expectations, the role of internationally traded financial assets, and the distinctions between stock and flow phenomena. It is to these theories that we now turn.

Exchange Rate Determination

A statement with which few economists would argue is that under flexible rates, the value of a currency is determined by supply and demand. Older theories, though, emphasized the supplies

and demands for foreign exchange arising from flow demands for merchandise imports and exports: the trade balance was seen as the major determinant of the exchange rate. These theories were consistent with the stylized facts of the time, namely relatively free trade in goods but restricted trade in financial assets. The overriding issue on which turned the workability of flexible rates was whether import and export elasticities were 'large enough' to ensure Walrasian stability of the foreign exchange market. Whether speculators were stabilizing or destabilizing in these theories depended mostly on what sort of expectations of future exchange rate movements speculators were assumed to have.

The international monetary theoreticians of the 1970s, in an effort to explain the unforeseen volatility of floating rates, began emphasizing the demand and supply for stocks of internationally traded financial assets. Monetary factors operating through interest differentials moved to the fore as the fundamental force behind exchange-rate determination, and the trade balance and relative prices of imports and exports were pushed to the background. These theories also incorporated Muth's ideas about rational expectations, setting the stage for analyses of speculation not critically dependent on ad hoc, arbitrary specifications of how expectations of future exchange rates are formed. The emphasis of these new theories on purchases on the capital account also reflected awareness that the 1970s were fundamentally different from earlier periods in that the world had fewer controls on capital movements.

These new asset-market theories of exchange-rate determination were refined and extended to incorporate trade-balance considerations, bringing relative prices back into the picture. What has now emerged is a theory with the following implications:

1. In the long run, a period measured in years rather than months or quarters, exchange rates are proportional to relative money supplies, so long as real factors remain roughly constant. That is, Cassel's Purchasing Power Parity Principle holds as a long-run phenomenon.
2. The short-run behaviour of exchange rates can be highly volatile, with exchange rates deviating markedly from their long-run trends. Speculators with rational expectations play an important role in this area by exacerbating monetary shocks to the system.

We now develop a skeletal model of exchange-rate determination that captures the key features of the asset-market approach. The key building blocks are specification of a stock demand for foreign financial assets, and specification of the supply of foreign assets. Expectations are modelled as rational, letting us focus on fundamental behavioural relations rather than on the effects of various ad hoc expectational schemes. The approach here differs from more standard treatments of asset market approaches, e.g. Frankel (1983), in its emphasis on how future trade-balance effects influence current exchange rates.

The hallmark of the asset-market approach is the specification of asset demands as stock, rather than flow, demands. Elementary mean-variance portfolio theory suggests, as a first approximation, that net demand for foreign financial assets should depend on relative rates of return. Denote the stock demand for net foreign assets as F_t, where t indexes time. Let e stand for the log of the exchange rate, defined as the domestic currency value of foreign exchange, and E_t denote the mathematical expectation of any variable conditional on information available at time t. The relative rate of return on foreign assets *vis-à-vis* domestic assets is approximately:

$$E_t e_{t+1} - e_t - r_t \tag{1}$$

where r_t is the difference between the domestic and foreign nominal interest rate. For simplicity, we specify the net demand for foreign assets as a linear function of the relative rate of return:

$$F_t = n\{E_t e_{t+1} - e_t - r_t\} \tag{2}$$

where n is a positive constant. Portfolio theory tells us that n should depend inversely on exchange rate predictability and directly on

taste for risk. If exchange rates become more unpredictable, then for any given expected return on foreign bonds the risk becomes larger. Hence, risk-averse investors will lower their holdings. Likewise, for any given expected return and degree of exchange-rate predictability, a decrease in risk-aversion by investors would lead them to increase their holdings of the risky asset.

Equilibrium in the foreign bond market means that demand equal supply. Denoting net foreign bond supply as F_s, this means that the exchange rate that equilibrates demand and supply of foreign assets satisfies the following equation:

$$e_t = E_{t+1}e_t - r_t - F_t^2/n. \tag{3}$$

That is, the current(log) exchange rate equals the current expectation of next period's rate minus the interest differential minus the stock of foreign bonds divided by the sensitivity of foreign bond demanders to expected returns. Note that if foreign bond demanders are risk neutral, i.e. if n is infinite, then the expected exchange rate change just equals the interest differential; all that investors care about is expected return, regardless of the relative riskiness of foreign bonds due to unforeseen exchange rate movements. This is the so-called efficient markets hypothesis for the foreign exchanges.

If we want to, we can think of e_t, the exchange rate, as moving to equilibrate demand and supply for net foreign bonds at each moment in time. Asset-market theorists have sometimes claimed that thinking of the exchange rate this way as opposed to thinking of it as moving to equilibrate flow supplies and demands for foreign exchange is what distinguishes the new approach from the old. Note, though, that a component of demand is E_te_{t+1}, the current expectation of next period's exchange rate. Now, if the exchange rate moves to equilibrate foreign net bond demand and supply at $t + 1$, then e_{t+1} depends on F_{t+1} and $E_{t+1}e_{t+2}$, which in turn depends on F_{t+2} and $E_{t+2}e_{t+3}$, which in turn depends on $F_{t+3}e_{t+4}$, and so on. Thus, the current exchange rate depends on current expectations of all future values of net foreign bond holdings.

We can make this argument formally by iterating equation (3) forward through time, taking expectations, and substituting back in the initial equation. We can then write the exchange rate as three terms: the current expectation of the long-run exchange rate, the current expectation of the sum of the current plus all future interest rate differentials, and the current expectation of the sum of current plus all future foreign bond supplies, divided by investor sensitivity to expected returns:

$$e_t = E_te_{t+1} - E_t\sum_{i=0}^{\infty} r_{t+1} - (1/n)E_t\sum_{i=0}^{\infty} F_{t+i}^s \tag{4}$$

What this equation highlights is how the entire future path of both interest rates and foreign bond supplies affects current exchange rates. If interest rate differentials are highly variable, perhaps reflecting variable monetary policy, then the exchange rate will be highly variable. The future values of foreign bonds highlight, indirectly, the role of the current account and relative prices in exchange rate determination.

Net additions to the stock of foreign bonds available to domestic residents can only be generated by trade balance surpluses. Symbolically, we have:

$$F_t^s - F_{t-1}^s = T_t \tag{5}$$

where T_t denotes the trade balance at time t. The simplest specification of the behaviour of the trade balance would make it depend upon relative prices. If we denote the log of relative price levels between domestic and foreign countries as p, then we can specify the trade balance as:

$$T_t = a\{e - p\} + u_t \tag{6}$$

where a is a parameter reflecting the responsiveness of the trade balance to relative prices, and u_t is a zero-mean serially uncorrelated random variable, capturing shocks to the underlying fundamental determinants of the trade balance, e.g. taste and technology.

At this point it is useful to develop the behaviour of the stock of foreign bonds through time.

For expositional ease, we assume that the interest differential, r_t, is a serially uncorrelated zero-mean random variable. It turns out that F_t will follow a first-order autoregressive process through time:

$$F_t = x_1 F_{t-1} + x_2 u_t + x_3 r_t \tag{7}$$

where x_1, x_2 and x_3 are coefficients which are functions of n and a. From the above process, it follows that

$$\begin{aligned} & E_t \sum_{t=0}^{\infty} F_{t+1}^s \\ & = F_t^s \left[1 + x_1^2 + \cdots + x_1^j + \cdots + x_1^j + \cdots \right] \\ & = F_t^s [1/(1 - x_1)] \end{aligned} \tag{8}$$

Hence, $et = E_t e_{t+1} - r_t - (1/n) F_t^s [1/(1 - x_1)]$. Assume for simplicity of exposition that relative price levels, p, are fixed. We could make any one of a variety of other standard macroeconomic assumptions about price-level determination without altering the basic lessons of this analysis. Then, $F_r = F_{t-1} + ae_t + u_t$. Using the immediately preceding expression for e_t in this equation, some simple algebra leads us to:

$$\begin{aligned} F_t = & \{[n(1 - x_1)]/[n(1 - x_1) + a]\} F_{t-1} + \\ & \{[n(1 - x_1)]/[n(1 - x_1) + a]\} u_t \\ & - \{[an(1 - x_1)]/[n(1 - x_1) + a]\} \end{aligned} \tag{9}$$

Comparing (7) and (9), we see that x_1 is implicitly defined by:

$$x_1 = \{[n(1 - x_1)]/[n(1 - x_1) + a]\} \tag{10}$$

and there exists one unique value of x_1 between zero and one. That is, F_t^t follows a stable autoregressive process whose parameters are functions of the two structural parameters n and a.

Armed with this knowledge, we can derive the path through time of the exchange rate. Differencing the fundamental equation (3) and substituting the trade balance for $(F_t - F_t - {}_1)$, we get the following ARMA(1, 1) process:

$$\begin{aligned} e_t = & x_1 e_{t-1} - x_1 r_t + x_1 r_t - 1 \\ & + \{[1 - x_1]/a\} u_t. \end{aligned} \tag{11}$$

Analysis of this equation shows us how the operating characteristics of flexible exchange rates are related to whether shocks to the system are 'real', i.e., u_t, or 'monetary', i.e., r_t, and to the 'aggressiveness' of speculators, that is, to the magnitude of n. Of course, to denote r_t as the 'monetary' shock ignores the influence of real factors on nominal interest rates. In a full general equilibrium model, r_t would represent a commingling of more fundamental real and monetary shocks. First note that x_1 is monotonically increasing between zero and one as n goes between zero and infinity. Hence, when speculators are risk-neutral (n infinite), the exchange rate is completely insulated from real shocks. On the other hand, the exchange rate is least insulated from monetary shocks in this case. Whether speculators stabilize or destabilize exchange rates is thus seen to be dependent on whether shocks are real or monetary. The surprise of economists over the variability of exchange rates in the 1970s can be thought of as their lack of appreciation of how international capital mobility transmits interest-rate disturbances internationally. Their realization now that high capital mobility and variable monetary policy can lead to volatile exchange rates has led some to call for 'throwing sand in the system'.

Finally note that the long-run properties of the above asset-market model correspond to purchasing power parity. In the long run, variables are at their steady states and stocks of foreign assets are no longer changing. The trade balance, then, must be zero. Hence, on average, the exchange rate must be equal to relative price levels. Relative price levels are themselves proportional to relative money supplies, real factors remaining constant. The long-run exchange rate then will be proportional to relative money supplies. The fact that for some periods and some exchange rates this prediction has been violated implies that permanent real shocks have been important sources of variability.

The following perspectives on flexible exchange rates emerge from the preceding analysis. First, an emphasis on the role of trade in international financial assets in exchange rate determination leads to an explanation of the volatility of exchange rates under a flexible rates regime. This does not negate, though, the ability of flexible rates to provide long-run monetary independence to individual countries; countries can choose different price levels in the long run, with the exchange rate moving to equilibrate.

Second, exchange rates are endogenous variables, and exchange rate volatility is in important ways a symptom of underlying volatility, not a cause. This volatility does have real effects, though, on relative international competitiveness and associated macroeconomic dislocations.

Finally, both real and monetary factors can play an important role in exchange rate determination. Coincident with this observation, we should note that our major interest is with real exchange rates, i.e. relative prices. In the model used in this entry, our assumption of fixed relative price levels let us identify the nominal exchange rate with relative prices. Of course, in reality price levels are not fixed. Even so, we observe in the world that real and nominal exchange rate movements are highly positively correlated, probably reflecting some sort of price-level stickiness, in which case our analysis is still relevant.

See Also

- ▶ Exchange Rates
- ▶ Fixed Exchange Rates
- ▶ International Finance
- ▶ Purchasing Power Parity

Bibliography

Frankel, J.A. 1983. Monetary and portfolio balance models of exchange rate determination. In *Economic interdependence and flexible exchange rates*, ed. J. Bhandari and B. Putnam, 84–115. Cambridge: MIT Press.

Friedman, M. 1953. The case for flexible exchange rates. In *Essays in positive economics*, ed. M. Friedman, 157–203. Chicago: University of Chicago Press.

Florence, Philip Sargant (1890–1982)

Murray Milgate

An institutional economist whose work fell broadly into the fields of regional development and industrial organization, Philip Sargant Florence was born in New Jersey (USA) on 25 June 1890 but lived and worked for the greater part of his life in England. He was educated at Rugby, Caius College, Cambridge (taking a First in the Economics Tripos in 1914), and Columbia University (Ph. D.). From 1921 until 1929 he was a lecturer at Cambridge, and was then elected (succeeding J.F. Rees) into the Chair of Commerce in the University of Birmingham, which he occupied until his retirement in 1955. As an American citizen he travelled back to the US for appointments as visiting professor after his retirement, and this he did on at least two occasions – once visiting Johns Hopkins and on another occasion the University of Rhode Island. For the last ten years of his life he was a vice-president of the Royal Economic Society. He died on 29 January 1982 at the age of 91.

Sargant Florence's work on practical problems of industrial organization seems to have developed out of a narrower concern in his earliest writings with the internal organization of productive activities and, in particular, with the effects of fatigue (together with other sociological factors) on the productive efficiency of labour (see, for example, 1918). For Sargant Florence, the problems of industrial organization required the consideration of many issues: market structure and regulation, the direct organization of production and its management, the length of the working day, the role of competition versus combination, the education of businessmen, and the efficient design of incentive systems and job ladders to name but a few (see, e.g., 1933). It is little wonder that after a lifetime of work on this kaleidoscope of questions he began to lean more and more towards the belief that the efficient organization of industry was a question that economics could only begin to answer with the help of sociology

(see, for example, 1964). It should be said, however, that some of Sargant Florence's writings on these difficult subjects seem to be little more than the reflections of an educated individual with an active but not very profound interest in society, rather than the considered conclusions of a careful social scientist.

Sargant Florence's interest in regional economics seems to have taken root soon after World War II when, on returning to England from the US where he had spent the better part of the war years, he resumed his academic duties at Birmingham and began active applied research into the problems and priorities of regional development and planning in the Midlands as part of post-war reconstruction initiatives in Britain. These investigations generated a number of publications in which his role as author, editor and contributor varied: *County Town* (1946) was a study of Worcester, *English County* (1947) was an examination of the industrial infrastructure of Herefordshire, and *Conurbation* (n.d.) delineated the problems confronting attempts at regional and urban planning in Birmingham and the Black Country.

It is perhaps worth recording that Sargent Florence also wrote a 500-page treatise on the statistical method in economics and political science in 1929. He dedicated this book to all those 'who find theories unsatisfactory without the test of fact' (p. v), and in the preface to the same thanks a number of now well-known Cambridge economists for advice (including Maurice Dobb, Joan Robinson, Gerald Shove, Austin Robinson and Lavington). However, Florence's idiosyncratic understanding of the relationship between theoretical argument and empirical evidence in this work is well illustrated by his criticism and attempted refutation of Freud's theory of parapraxis (1929, pp. 196–9). Towards the end of his life, in 1975, Sargant Florence rather uncharacteristically took up the macroeconomic question of the causes of inflation.

Selected Works

1924. *The economics of fatigue and unrest*. London/New York: G. Allen & Unwin/H. Holt & Co.
1926. Over-population: Theory and statistics. London: Kegan Paul & Co.
1927. *Economics and human behaviour*. London/New York: Kegan Paul & Co./W.W. Norton.
1929. *The statistical method in economics and political science*. London/New York: Kegan Paul, Trench and Trubner/Harcourt, Brace & Co.
1930. *Uplift in economics*. London: Kegan Paul & Co.
1933. *The logic of industrial organization*. London: Kegan Paul, Trench & Trubner.
1938. (With A. Carr-Saunders et al.) *Consumers' co-operation in Great Britain*. New York/London: Harper & Brothers.
1948. *Investment, location, and size of plant*. Cambridge: Cambridge University Press.
1953. *The logic of British and American industry*. London/Chapel Hill: Routledge & Kegan Paul/University of North Carolina Press. Revised edn, 1971.
1957. *Industry and the state*. London: Hutchinson's University Library.
1964. *The economics and sociology of industry*. London: C.A. Watts & Co. Revised edn, 1969.
1975. Stagflation in Great Britain: The role of labor. in *The roots of inflation*, ed. Gardiner C. Means et al. New York: Burt Franklin & Co.

Flux, Alfred William (1867–1942)

J. K. Whitaker

A distinguished applied economist and statistician, Flux was born in Portsmouth on 8 April 1867, the son of a journeyman cement maker. He died in Denmark, his wife's native land, on 16 July 1942. After entering St John's College, Cambridge, he was bracketed as Senior Wrangler in the Mathematics Tripos of 1887. Soon turning to economics, he came under Alfred Marshall's influence, joining Marshall as a Fellow of St John's in 1889. Leaving Cambridge in 1893 to teach economics at Owens College, Manchester,

he next moved in 1901 to McGill University, Montreal. In 1908 he returned to London as statistical adviser to the Board of Trade, where he remained until retirement in 1932, being knighted in 1934.

To the pre-1908 phase belong: a steady stream of pioneering statistical studies of international trade exemplified by Flux (1894b, 1897, 1899); the first post-Marshallian British textbook (Flux 1904), an accomplished but unoriginal exposition with an interesting geometrical appendix; a new edition of Jevon's *Coal Question* (Flux 1906); and, though it appeared only after 1908, a study of Swedish banking for the US National Monetary Commission (Flux 1910). But the work for which he is now best known, his only significant theoretical contribution, was his earliest publication, a review of Wicksteed's *Coordination* (Flux 1894a) which first invoked the Euler theorem to prove that marginal productivity imputation just exhausts output given constant returns to scale in production.

After 1908, Flux found his *métier* in the development of official statistics. A series of papers given to the Royal Statistical Society, exemplified by Flux (1913, 1921, 1927, 1929), stands as monument to his important contributions. See also his Newmarch Lectures (Flux 1924).

Flux remained a frequent reviewer for the *Economic Journal* but never matched his first performance. He also contributed to the original *Palgrave*. No comprehensive bibliography has been compiled of his many articles and pamphlets. But only three are of a theoretical or doctrinal character, none of these being especially noteworthy. For biographical details see Chapman (1942).

Selected Works

1894a. Review of Wicksell, K., Kapital und Rente nach der Neuern Nationalökonomischen Theorien, and Wicksteed, P.H., Essay on the coordination of the laws of production and distribution. *Economic Journal* 4: 305–313.

1894b. The commercial supremacy of Great Britain. *Economic Journal* 4(Pt. I): 457–467; Pt. II: 595–605.

1897. British trade and German competition. *Economic Journal* 7: 34–45.

1899. The flag and trade: A summary review of the trade of the chief colonial empires. *Journal of the Royal Statistical Society* 62: 489–522.

1904. *Economic principles*, 2nd ed. London: Methuen, 1923.

1906. In *The coal question*, 2nd ed., ed. W.-S. Jevons. London: Macmillan.

1910. *The Swedish banking system*. Publications of the National Monetary Commission, Vol. XVII, No. 1. Washington, DC: Government Printing Office.

1913. Gleanings from the census of production report. *Journal of The Royal Statistical Society* 76: 557–585.

1921. The measurement of price changes. *Journal of the Royal Statistical Society* 84: 167–199.

1924. *The foreign exchanges*. Newmarch Lectures for 1922. London: P.S. King.

1927. Indices of industrial productive activity. *Journal of the Royal Statistical Society* 90(2): 226–258.

1929. The national income. *Journal of the Royal Statistical Society* 92(1): 1–25.

References

Chapman, S.J. 1942. Sir Alfred William Flux. *Economic Journal* 52: 400–403.

Flypaper Effect

Robert P. Inman

Abstract

The flypaper effect results when a dollar of exogenous grants-in-aid leads to significantly greater public spending than an equivalent dollar of citizen income: money sticks where it hits. Viewing governments as agents for a representative citizen voter, this empirical result is an anomaly. Four alternative explanations have

been offered. First, it is a data problem; exogenous aid is mismeasured. Second, it is an econometric problem; important explanators of spending correlated with aid or income are excluded from the specification. Third, it is a specification problem; the representative citizen misperceives aid and the rational voter model misses this point. The empirical evidence suggests none of these explanations is sufficient. A fourth explanation seems most promising: it is politics. Rather than an anomaly, the flypaper effect is best seen as an outcome of political institutions and the associated incentives of elected officials.

Keywords

Flypaper effect; Grant aid; Political spending; Public funds

JEL Classifications

H72; H77; P16

In the late 1960s James Henderson (1968) and Edward Gramlich (1969) changed the direction of empirical research on how local governments tax and spend. While all prior work detailed the demographic and economic correlates with government budgets, Henderson and Gramlich sought an explanation for those correlations. To them as economists, the answer was clear. Citizens demand services from their elected officials, and elected officials respond subject to the availability of government resources. Resources come from citizen incomes and from fiscal transfers given by the central government as grants-in-aid. From this perspective, Henderson and Gramlich specified and estimated demand equations based on the maximization of a representative citizen's utility subject to that citizen's 'full income' constraint specified as the sum of personal income and the citizen's share of the government's unconstrained fiscal transfers. So specified, personal income and the citizen's share of fiscal transfers should impact spending identically – money is money.

The empirical analyses of Henderson and Gramlich revealed something unexpected, however. An extra dollar of personal income increased government spending on the order of \$0.02 to \$0.05, but an equivalent extra dollar of grants-in-aid increased government spending by from \$0.30 to often as much as a full dollar. When Gramlich first presented his results, his colleague Arthur Okun called this larger effect of lump-sum aid on government spending a 'flypaper effect', noting that 'money seems to stick where it hits'. The label stuck too, as has the puzzle of why intergovernmental transfers are so stimulative. A Google search reveals that over 3,500 research papers – excluding those studying the effects of real flypaper on insect populations – have now been written documenting and seeking to explain the flypaper effect.

Why do we care about this apparent anomaly? There are two reasons. First, as a matter of policy, understanding *how* recipient governments spend intergovernmental transfers is essential for the design of efficient fiscal policy in federal economies. Second, as a matter of science, understanding *why* governments spend citizens' incomes as they do provides valuable insights into how citizen preferences are represented in government policies. The taxation of citizen incomes and the allocation of grants-in-aid provide two 'tracers' as to the inner workings of political decision-making, one (taxes) that is directly observed and controlled by citizens, and the other (grants) perhaps only imperfectly so.

The benchmark for both the policy and political economy literatures is how a politically decisive citizen would like to see government resources allocated, specified by the maximization of that representative citizen's welfare over private (x) and public (g) goods, indexed by U(x, g), subject to a current period budget constraint specified as:

$$Y = \{I + h \cdot z\} = x + p_g \cdot g$$

where I is the citizen's private income (or tax base), h is the citizen's share of *unconstrained or lump-sum* intergovernmental transfers per capita (z) specified as $h = I/\bar{I}$ with $\bar{I}$ equal to the average income (or tax base) in the citizen's political jurisdiction, and p_g is the 'tax price' for government

services (g) equal to c (1 − m) h where c is the per unit production cost of g and m is the matching rate for openended matching federal aid. Private goods cost $1. Y is called the citizen's 'full income'. The citizen's preferred allocations will be $x = x(1, p_g, Y)$ and $g = g(1, p_g, Y)$, where: $\Delta g_I = (\delta g/\delta Y) \cdot (\delta Y/\delta I) \cdot \Delta I = (\delta g/\delta Y) \cdot (\Delta I = \$1)$, for an extra dollar of personal income and: $\Delta g_z = (\delta g/\delta Y) \cdot (\delta Y/\delta z) \cdot \Delta z = (\delta g/\delta Y) \cdot h \cdot (\Delta z = \$1)$ for an extra dollar of aid, implying that estimated marginal effects of aid to income should be related as $\Delta g_z/\Delta g_I = h$. In most political jurisdictions the representative citizen has a tax base (often specified as the median tax base) less than the average tax base; thus, in most cases, if our representative citizen has had her way, then we should expect $\Delta g_z/\Delta g_I = h < 1$. The overwhelming empirical evidence summarized by Gramlich (1977), Inman (1979), Fisher (1982) and Hines and Thaler (1995) shows just the opposite, however; Δg_I ranges from $0.02 to $0.05 while the companion estimates of Δg_z typically fall between $0.30 and $1.00. Income to the citizen stays with the citizen; grants to the government stay with the government.

Money sticks where it hits. Why?

Four explanations have been offered. First, the answer is in the data. Researchers mismeasure intergovernmental aid by confusing matching grants that lower the marginal price of public services (p_g) with lump-sum aid (z) that shifts outward the representative citizen's budget constraint. Matching aid has a price effect, lump-sum aid an income effect. For local politics controlled by a representative citizen, consumer theory predicts that a matching grant's price effect will stimulate more government services than an equivalent dollar of lump-sum aid. If the dollar transfer received from matching aid is erroneously classified as lumpsum aid, then $\Delta g_z > \Delta g_I$ will result; see Moffitt (1984), Megdal (1987), and Baker et al. (1999). Even after correctly classifying aid programmes and measuring p_g and z appropriately, however, the flypaper effect remains; see for example Wyckoff (1991).

The second explanation sees the anomaly as an econometric problem. Researchers may have omitted important determinants of government spending likely to be correlated with citizen income or intergovernmental aid, leading to biased estimates of Δg_I and Δg_z. Bruce Hamilton (1983) and Jonathan Hamilton (1986) attribute the flypaper effect to misspecifications of the technology or costs of providing local services. Bruce Hamilton argues that estimated demand equations omit important variables such as the citizen's talents or willingness to volunteer which are positively correlated with citizen income and also contribute to the provision of government services. If these omitted effects are substitutes for (negatively correlated with) purchased government inputs, then the estimated coefficient for income will be biased downward, perhaps sufficiently so that $\Delta g_z > \Delta g_I$. Jonathan Hamilton suggests the misspecification arises from a failure to account correctly for residential exit from high tax jurisdictions leading to a loss of tax base when specifying the price of government services. Local taxes are inefficient and the correctly specified price of local services must reflect this fact. If citizens tend to reside in localities of comparable income, and higher-income residents are more mobile, then the representative citizen's income will be positively correlated with the correct price, which is negatively correlated with government services. Again, there is a downward bias in the estimated income effect, with $\Delta g_z > \Delta g_I$ as a possible result.

Neither of the Hamiltons's biases are likely to fully explain estimated flypaper effects, however. A plausible upper estimate for Δg_I can be obtained as $\Delta g_I = (\delta g/\delta Y) = \varepsilon_{g,Y} \cdot (g/Y)$, where $\varepsilon_{g,Y}$ is the income elasticity of demand for government services and g/Y is the average rate of spending by recipient governments. This ratio for the US state and local government sectors combined from 1970 to 2008 – the period used for most all studies – is at most 0.15. Since most state and local services are arguably necessities, $\varepsilon_{g,Y} \leq 1$ seems reasonable. If so, then $\Delta g_I \leq (g/Y) = 0.15$ bounds an unbiased income effect. Since most estimates of Δg_z exceed 0.15, the flypaper effect remains.

Perhaps then the explanation lies in an upward bias in the estimates of Δg_z? Here the results of four recent studies are particularly instructive.

Each takes advantage of a plausibly exogenous, or 'natural experiment', change in lump-sum national aid to state or local governments. Gordon (2004) uses US federal legislation's required changes in Title I education aid caused by state-level (exogenous to the local budget) demographic changes before and after census years as her measure of exogenous aid. She finds strong evidence of a flypaper effect for local school districts in the first year after the change in Title I aid – $\Delta g_z = 1.00$ – but that this effect evaporates after three years, with most of the new aid returned to voters as lower local tax revenues. In contrast, Ladd (1993) and Singhal (2008) find evidence for a significant and quantitatively large flypaper effect for US state governments, as do Dahlberg et al. (2008) in their study of national aid to municipalities in Sweden. Ladd uses windfall tax revenues to state governments following the Tax Reform Act of 1986 as her exogenous measure of aid, and estimates $\Delta g_z = 0.40 > \Delta g_I = 0.03$. Singhal (2008) uses outside revenues received by state governments from a recent legal settlement with the tobacco industry as her measure of z, and finds $\Delta g_z = 0.20$ for spending on tobacco control programmes, compared with an estimate of $\Delta g_I \approx 0$ for income's effects on the same programmes. Dahlberg et al. (2008) exploit a discontinuity in the national aid formula that gives significant additional assistance to communities that experience more than 2 per cent outmigration over the previous ten years; communities just below the threshold receive no additional aid, those just above do. The analysis includes community and time fixed effects – there is no direct estimate of Δg_I – and they find $\Delta g_z = 1.00$ and no local tax relief. Ladd's, Singhal's and Dahlberg's estimated flypaper effects remain over time.

The flypaper effect appears to be a real phenomenon. As a third explanation, then, perhaps our model of citizen fiscal choice is misspecified. First, voters may not understand the complexity of grant programmes. Both Courant et al. (1979) and Oates (1979) conjecture that the representative citizen misperceives lump-sum aid's income effect as an average price effect. They conjecture that the voter uses taxes paid per unit of services received – $(p_g \cdot g - z)/g$ or $p_g - (z/g)$ – as their estimate of the true marginal tax cost of government services, p_g. If so, lump-sum aid (z) will impact spending as a price subsidy, and the estimated effect aid on spending will imply that $\Delta g_z > \Delta g_I$. Wyckoff (1991) and Turnbull (1998) test this hypothesis by including both p_g and $[p_g - (z/g)]$ as competing explanators of local spending. They find plausible (negative) marginal price effects but implausible (positive) effects of the misperceived average price. Estimated flypaper effects are comparable to those of previous studies. From this evidence, it is unlikely that price misperception provides the explanation for the flypaper effect.

Filimon et al. (1982) and Hines and Thaler (1995) provide alternative versions of the voter ignorance hypothesis. For Filimon, Romer and Rosenthal the representative voter fails to see through the veil of government budgets; he does not know the level of aid received by the local government. For Hines and Thaler, the representative voter sees through the veil but budgets using mental accounts; there is a 'public budget' that is the responsibility of government officials and a 'private budget' that is the citizen's responsibility. Both hypotheses need a theory of public budgets to explain Δg_z. Hines and Thaler leave this an open question, but Filimon, Romer, and Rosenthal are quite explicit: public officials are budget maximizers and therefore $\Delta g_z = 1$. They test their theory for a sample of Oregon school districts, and cannot reject the null hypothesis that $\Delta g_z = 1$ for state education aid.

In Romer et al. (1992), the authors replicate their analysis for a sample of New York school districts, and here the conclusion varies by the size of the school district. Large districts (>20,000 students) show budget-maximizing behaviour and a full flypaper effect: $\Delta g_z = 1$. In smaller districts, however, the estimated aid and income effects are about equal: $\Delta g_z \approx h \cdot \Delta g_I$ These results parallel those from Ladd and Singhal for larger state governments and from Gordon for local school districts. Together, this evidence is sufficient to reject a strict version of the mental accounting explanation. It leaves open, however, the question of why the flypaper effect remains for larger governments.

Here a fourth explanation for the flypaper effect seems the most promising: it is politics. This approach assumes that voters are informed and rational, but conceal their preferences when it is strategically useful to do so. Such strategic behaviours require the use of less than efficient institutions for preference revelation, such as majority rule or representative legislatures. From this perspective, the flypaper effect is a consequence of an inability of citizens to write complete 'political contracts' with their elected officials. Consistent with the results of Ladd, Singhal, and Romer, Rosenthal and Munley, we might expect these contracting problems to be greater, and the flypaper effect more likely, for large governments.

Chernick (1979) and Knight (2002) offer specifications of a political contract between a donor central government and recipient local governments as a way to understand the flypaper effect. Chernick (1979) specifies donor-recipient contracting as an auction. Assuming an exogenous level of federal aid, local governments bid for the right to provide aided services by offering to share the costs of provision. Beginning with the highest offer price, the central government selects recipient local governments until its grants budget is exhausted. The resulting allocation will equalize the marginal contribution of each local government to the incremental benefits from the provision of the local service. Local governments with the highest valuations will provide more services and receive more aid. Chernick offers evidence in support of this prediction from the US federal Water and Sewer Grant programme. Importantly, any reduced form estimate of Δg_z for this programme that did not account for the auction that sets aid would be biased upward and imply a strong flypaper effect.

Knight (2002) specifies and estimates a model of political contracting for grants policy that sets both the aggregate size of the aid budget and its allocation. The budget is chosen to ensure its passage and to maximize local constituent net benefits for the central government's agenda-setter. Again, the allocation process is an auction. Legislators bid to be part of the winning coalition by offering to vote for the grants budget in return for intergovernmental aid. The agenda-setter picks the smallest 51 per cent of the bids. He then sets his own grant award to maximize the net benefits to his own constituents. Those legislators whose state or local governments value the aided local service most highly make the winning offers. The result is again a positive correlation between grants awarded and local spending. Failure to control for this correlation will lead to an upward bias in the estimate of Δg_z. For a statistically consistent estimate of Δg_z we need instruments that both predict grants (z) and are independent of constituents' demand for the aided service. Legislative institutions that select agenda-setters independent of constituent preferences will serve this purpose. Knight uses the legislators' tenures and majority party memberships as his instruments in his empirical study of highway grants and state highway spending. Least squares estimation of grants' effect on spending shows $\Delta g_z = 1$; instrumental variables estimation rejects that extreme flypaper result but cannot reject a partial effect ($1 > \Delta g_z > h \cdot \Delta g_I$). In a companion piece, Knight (2004) estimates that this agenda-setting process for highway grants imposes an allocative inefficiency of $0.40 per dollar of aid.

Over the first decade of the 21st century, the devolution of economic responsibilities to lower-tier governments has become increasingly important, not only in formally federal states but in unitary states as well. Central governments typically grant fiscal assistance to these local governments for the provision of those services. Knowing how grants will be spent is important for the appropriate design of central government transfer policies. Credible estimates of aid's effects on local spending requires good instrumental variables to predict aid, or ideally 'natural experiments' providing truly exogenous measures of central government assistance. Knowing how money is spent as it is helps us to understand allocative performance of intergovernmental transfers, given federal and local political institutions.

Knowing why grant money is spent as it is, is just as important. Here the specification and estimation of structural models of central government transfer spending and local government

allocations of transfer incomes are essential. This information provides a basis for reforming these important institutions, and there is perhaps no more striking example of the benefits of such structural analyses of the aid process than the work of Reinikka and Svensson (2003, 2004) on the allocation of Ugandan central government aid to local schools. Initially, only $0.15 of each centrally allocated school aid dollar found its way into the local schools; $0.85 was 'captured' by the district bureaucracy for its own use. The problem was inadequate information and weak local political organizations. Reforms publicized aid allocations and empowered village councils to monitor that spending. The end results was to reduce district capture to $0.15 per aid dollar – a plausible administrative cost – and to increase local school resources by $0.85 per aid dollar.

Once viewed as an anomaly, the flypaper effect should now be seen as a reality of fiscal politics, and its study as an opportunity to fashion central government transfer policies and intergovernmental fiscal institutions that better reflect citizen preferences for local public goods.

See Also

- ▶ Foreign Aid
- ▶ Intergovernmental Grants
- ▶ Political Institutions, Economic Approaches to

Bibliography

Baker, M., A. Payne, and M. Smart. 1999. An empirical study of matching grants: The 'cap' on CAP. *Journal of Public Economics* 72: 269–288.

Chernick, H. 1979. An economic model of the distribution of project grants. In *Fiscal federalism and grants-in-aid*, ed. P. Mieszkowski and W. Oakland. Washington, DC: Urban Institute Press.

Courant, P., E. Gramlich, and D. Rubinfeld. 1979. The stimulative effects of intergovernmental grants: Or why money sticks where it hits. In *Fiscal federalism and grants-in-aid*, ed. P. Mieszkowski and W. Oakland. Washington, DC: Urban Institute Press.

Dahlberg, M., E. Mörk, J. Rattsø, and H. Ågren. 2008. Using a discontinuous grant rule to identify the effect of grants on local taxes and spending. *Journal of Public Economics* 92: 2320–2335.

Filimon, R., T. Romer, and H. Howard Rosenthal. 1982. Asymmetric information and agenda control. *Journal of Public Economics* 17: 51–70.

Fisher, R. 1982. Income and grant effects on local expenditures: The flypaper effect and other difficulties. *Journal of Urban Economics* 12: 324–345.

Gordon, N. 2004. Do federal grants boost school spending? Evidence from Title I. *Journal of Public Economics* 88: 1771–1792.

Gramlich, E. 1969. State and local governments and their budget constraint. *International Economic Review* 10: 163–182.

Gramlich, E. 1977. Intergovernmental grants: A review of the empirical literature. In *The political economy of federalism*, ed. W.E. Oates. Lexington: Lexington Books.

Hamilton, B. 1983. The flypaper effect and other anomalies. *Journal of Public Economics* 22: 347–362.

Hamilton, J. 1986. The flypaper effect and the deadweight loss from taxation. *Journal of Urban Economics* 19: 148–155.

Henderson, J. 1968. Local government expenditures: A social welfare analysis. *Review of Economics and Statistics* 50: 156–163.

Hines, J., and R. Thaler. 1995. Anomalies: The flypaper effect. *Journal of Economic Perspectives* 9: 217–226.

Inman, R. 1979. The fiscal performance of local governments: An interpretative review. In *Current issues in urban economics*, ed. P. Mieszkowski and M. Straszheim. Baltimore: Johns Hopkins Press.

Knight, B. 2002. Endogenous federal grants and crowd-out of state government spending: Theory and evidence from the Federal Highway Aid Program. *American Economic Review* 92: 71–92.

Knight, B. 2004. Parochial interests and the centralized provision of local public goods: Evidence from congressional voting on transportation projects. *Journal of Public Economics* 88: 845–866.

Ladd, H. 1993. State responses to the TRA86 revenue windfalls: A new test of the flypaper effect. *Journal of Policy Analysis and Management* 12: 82–104.

Megdal, S. 1987. The flypaper effect revisited: An econometric explanation. *Review of Economics and Statistics* 59: 347–351.

Moffitt, R. 1984. The effects of grants-in-aid on state and local government spending: The case of AFDC. *Journal of Public Economics* 23: 279–305.

Oates, W. 1979. Lump-sum intergovernmental grants have price effects. In *Fiscal federalism and grants-in-aid*, ed. P. Mieszkowski and W. Oakland. Washington, DC: Urban Institute Press.

Reinikka, R., and J. Svensson. 2003. *The power of information: Evidence from a newspaper campaign to reduce capture*. World Bank: Mimeo December.

Reinikka, R., and J. Svensson. 2004. Local capture: Evidence from a central government transfer in Uganda. *Quarterly Journal of Economics* 93: 6562–6587.

Romer, T., H. Rosenthal, and V. Munley. 1992. Economic incentives and political institutions: Spending and

voting in school budget referenda. *Journal of Public Economics* 4: 1–33.
Singhal, M. 2008. Special interest groups and the allocation of public funds. *Journal of Public Economics* 92: 548–654.
Turnbull, G. 1998. The overspending and flypaper effects of fiscal illusion: Theory and empirical evidence. *Journal of Urban Economics* 44: 1–26.
Wyckoff, P. 1991. The elusive flypaper effect. *Journal of Urban Economics* 30: 310–328.

Fogel, Robert William (Born 1926)

Stanley L. Engerman

Abstract

A pioneer in the development of cliometrics, Robert Fogel has always focused on linking economic analysis to the study of historical problems and on the need for large-scale data collection and analysis. Based on this approach, he found that that slavery was profitable and viable even on the eve of the American Civil War; and he used information on height to draw inferences on food consumption by slaves. Fogel has since become involved in the economics of aging and longevity, the impact of the expansion of leisure time in the developed world, and the increasing burden of health care.

Keywords

Aging populations; Agriculture; Anthropometric history; Cliometrics; Counterfactuals; Economic history; Fogel, R; Health care; Nutrition; Railroads; Slavery; Transport

JEL Classification

B31

Robert William Fogel is one of the pioneering figures in the development of cliometrics or the new economic history during the 1950s, a contribution for which he was awarded, with Douglass C. North, another innovator, the 1993 Nobel Prize in economic science. As of 2005 they are the only two economic historians to obtain this honor. Fogel was born in New York City on 1 July 1926, and graduated from Cornell University in 1948. Active politically in left-wing organizations for several years, he did not begin graduate work in economics until after 1956. He received a master's degree from Columbia University writing under the supervision of Carter Goodrich. He then went to the John Hopkins University, receiving a Ph.D. under the direction of Simon Kuznets in 1963. He has held teaching positions at the University of Rochester, the University of Chicago, and Harvard University, being the Charles R. Walgreen Distinguished Professor of American Institutions at the University of Chicago since 1981.

Fogel's career has always focused on the linking of economic analysis to the study of historical problems. The application of economic theory to specific historical questions has characterized his writings since he began graduate work. Also characteristic of his work has been a concern with empirical data, at first the use of quantitative data to study specific problems, and later, with developments in computer technology, with attention given to the collection and analysis of large data-sets of economic and demographic evidence.

Railroads

Fogel's first book, *The Union Pacific Railroad: A Case in Premature Enterprise*, based on his master's thesis, was published in 1960. The basic question asked was whether the building of the Union Pacific in the 1860s, with government subsidy, led to corruption and abnormal profits earned by the railroad's builders and promoters. It had long been a staple of historical scholarship about the Union Pacific that the charges of corruption were true, and that the Union Pacific was to be viewed as part of America's late nineteenth-century 'Great Barbecue' and the 'Gilded Age'. Fogel's extensive primary research permitted some more accurate accounting measures of the profits, and he used data on bond prices and related information to estimate the anticipated

'risk of failure' at the time of financing. By adjusting the accounting profits and pointing to the great measured risk in this pioneering transcontinental venture, Fogel argued that the extent of abnormal profits was overstated, but also that the mixture of public and private financing may not have been the most effective way to undertake construction. The novel historical application of economic theory here was in the measuring of the market assessment of risk on the basis of standard financial models.

Fogel's next book, *Railroads and American Economic Growth: Essays in Econometric History*, was based upon his doctoral dissertation. Published in 1964, it has become one the two early classics of cliometrics, the other being the study of the economics of American slavery by Alfred H. Conrad and John R. Meyer (1958). This work, aimed at estimating the contribution of the railroad to American economic growth in the nineteenth century, something that many contemporaries had discussed, led to significant debates about both economic techniques of measurement and the methodological principles of historical analysis. Fogel's book (and that published a few years later by Albert Fishlow 1965) asked, on the basis of considerable empirical information, what the estimated difference in costs was between shipping goods by railroad and shipping them by the next-best alternative: road, canal, river, or lake. This was used to measure what Fogel called the 'social savings' based on the difference in costs of shipping between railroads and alternatives, and was to be the basic measure of the railroads' contribution to economic growth in 1890. That the number came out smaller than expected led to some critiques of the analysis, but in a controversial next step Fogel argued that this was too high, since it did not allow for possible structural adjustments in the economy to the absence of a railroad, including the building up of a canal network that never existed, but seemed feasible, if necessary, and for which ample amounts of water existed. Also, following the development economics of the period, Fogel estimated the contribution of the railroad via backward linkages (for inputs) and forward linkages (for outputs). In general none made the contribution of the railroad as large as expected, a point that has been used to argue that no single innovation can itself explain much growth, and that for an economy to be successful there has to be a broad spread of productivity gains within the economy. Whatever the specific criticisms, the overall fruitfulness of Fogel's method of analysis is seen in the number of country studies undertaken using his approach for the study of the economic effects of the railroad. Some studies indicate a small growth contribution, although in several cases (particularly Mexico) the measured effects were large due to the lack of good alternative means of transportation.

A major debate concerning the use of so-called 'what if' or counterfactual history in Fogel's analysis arose primarily among historians. To economists used to drawing supply and demand curves and discussing the impact of changes, this approach was rather standard and not questioned, and one might have felt that, given the form of most historical analysis, the same general acceptance would be expected among historians. This was not, however, a general view, and the explicit use of counterfactuals led to much debate. In some cases there was, as there should be, a questioning of the appropriateness of the particular counterfactuals used, since, as argued later, a counterfactual based on Napoleon using an atomic bomb is of doubtful usefulness. In other cases, however, the criticism was of the general use of the approach, with the implications that no 'what if' statement can be used at any time. This debate has disappeared in recent years, with apparent agreement that counterfactuals have long been part of the historian's approach to the past, and their use is a generally accepted, if not necessary, part of any historical study.

Economics of Slavery

Fogel's next major project concerned one of the major issues of American historiography, the economics of slavery in the United States South. This project led to numerous publications, including several books and many articles, over a 30-year period by Fogel, his colleagues, and his students.

The first major publication, in 1974, was the two-volume *Time on the Cross* (co-authored with Stanley Engerman); the first volume subtitled *The Economics of American Negro Slavery*, aimed at a large audience, and the second subtitled *Evidence and Methods: A Supplement*, which contained more detailed descriptions of data and analysis, aimed primarily at a professional, scholarly readership. These works presented findings from numerous types of primary data located in southern archives as well as census publications and manuscripts, and used many research assistants to collect and analyse the primary data, a practice then not typical in either history or economics. Earlier work on manuscript data from the federal census of 1860, prepared by William N. Parker (1970) and Robert E. Gallman (1970), was also of particular use, and the works of numerous historians and economists, in previous decades as well as that available from the then booming area of slavery studies, was important in shaping the arguments. Both because of its heavy use of quantitative methods and also because of several of its major findings that seemed to go against some then commonly held views, *Time on the Cross* attracted an unexpected amount of attention and criticism for an academic publication, and there emerged a rather extended series of debates on many of the questions studied, leading to the publication of several books and many articles developing these disagreements. As before, some of the debate was about the nature of questions asked and some about the specifics of the substantive analysis.

The major economic findings in *Time on the Cross* were that slavery was profitable and was expected, by southerners and others, to be viable even on the eve of the Civil War. These findings were based on standard measures of profitability and price–rental ratios, but the calculation required collections of data on slave prices (by age, sex, and so on), slave productivity, slave demography, and the material consumption allowed to the slaves by their masters. While profitability and viability had not always been widely accepted, by the time the debates ended they did seem to be acceptable, suggesting that slavery would not collapse under of its own weight, that southern planters had behaved in a manner that indicated a responsiveness to economic incentives, and, moreover, with the use of related evidence, that the South was doing quite well economically on the eve of the Civil War. Two other arguments were, and remain, still somewhat debated. A straightforward economist's measure of the relative productivity of northern and southern agriculture in 1860, used to answer a question long-discussed by contemporaries and many subsequent scholars, compared agricultural output with inputs of land, labour, and capital, and indicated that southern agriculture was more 'efficient' than northern agriculture and that within the South it was the larger slave-using plantations (over 16 slaves) that were more 'efficient' than were the small, free white farms and smaller slave farms. The concept of efficiency was interpreted by some, not as a standard concept of economic measurement, but as a measure with distinct moral overtones. The findings for the South led a discussion of economies of scale in slave plantations in the United States and in Caribbean sugar production, and the importance of scale has been seen to be significant for understanding slave societies as well as for evaluating the economic adjustment to the emancipation of slaves. A second continuing controversy concerned what was regarded as the favorable material treatment allowed slaves, based upon estimates of consumption allowed by masters, and the argued-for limited impact on slave family and cultural life. The former argument was based on demographic and related evidence. These questions have now become more important, and the ability of slaves to defeat masters' attempts to exercise complete power over slaves is now more widely argued for in slave studies. Nevertheless, some disagreements on these issues remain.

In the aftermath of the *Time on the Cross* debates various articles by its co-authors and others were written for conference presentation and for publication. In 1989 Fogel published a new book on slavery, *Without Consent or Contract: The Rise and Fall of American Slavery*, which covered some of the earlier themes but also provided much new information on the politics of abolition in Britain and the United States. In general Fogel expanded on several

discussions, particularly on cultural and demographic matters, but he still maintained most of the basic positions of his earlier writings on slavery, and the book is more of a defence than a revision of those arguments. In 1992 three edited volumes of earlier papers and notes by Fogel and others were published, adding greatly to the information and analysis of *Without Consent or Contract*. Fogel was invited to give the William Lynwood Fleming Lectures at Louisiana State University in 2001, published in 2003 as *The Slavery Debates, 1952–1990*. This brief, non-technical volume reviews the many debates on slavery, examines the trends and shifts in the study of slavery in the United States, and provides a very useful summary of changes in views over a 50-year period.

Heights and Demographic History

One of the debates concerning the material conditions of slave life related to the issue of food consumption and nutrition. Only after the publication of *Time on the Cross* did Fogel and his collaborators become aware of the valuable information provided by information on height. The collection of this data from coastal shipping manifests of slaves carried in the interstate movement between 1808 and 1865, and other sources such as military records and the registrations of free blacks, turned out to be exceptionally important, both for comparisons of the heights of southern slaves with other populations, supporting the argument of basically adequate consumption by slaves, and in opening another major project for Fogel and for other economic historians, historians, and economists. There were, of course, some difficulties in making inferences about food consumption from information on height, given differences in work regimen and disease environments and some truncations introduced by height requirements.

Nevertheless, so widespread were available data on heights in many countries over long periods of time, mainly from military records, that the study of height and its use as an alternative (or complementary) measure of welfare became widely used by economic historians in many different countries. Studies by Floud et al. (1990), Komlos (1994), Steckel (1995), Goldin and Rockoff (1992), and Steckel and Floud (1997), among others, frequently utilized measures for comparative purposes across countries, as well as for studying long-term trends within specific countries. Some unexpected patterns developed, such as long-period cycles in height, rather than simply monotonic change, and periods of time in which heights and per capita incomes move in different directions. Fogel used the study of heights as a method of approaching a number of different problems, such as long-term variations in longevity and health and their contributions to economic growth, changes in diseases and patterns of aging, and the economics of the health care industry. As earlier, several of these projects were based on extensive data retired from archival sources, and required collaborative work with many scholars from different disciplines.

In 1996 Fogel gave the McArthur Lectures at Cambridge University and these were published in 2004 as *The Escape from Hunger and Premature Death, 1700–2100: Europe, America, and the Third World*. In these essays he was concerned with changes in productivity due to improvements in human capacity to perform. He focused on changes in the twentieth century in health, the dramatic change in the caloric input of the French and British populations from their earlier limited available energy, and also the great increases in available leisure time. Such benefits of health and leisure have not yet occurred in much of the Third World today, where people adapt to limited energy by a smaller body size, limiting the prospective productivity in these societies relative to that in the developed world.

Fogel's most recent project, based on extensive data collection from archival sources and involving many students and scholars in collection and analysis, concerns long-term longitudinal studies of health, diseases, and the role of socio-economic and biomedical factors. The initial major data-set was based on the pension records of the Union army in the Civil War, which present very detailed medical histories of veterans from childhood until death, and run from the Civil War into the twentieth century. These data, with more recent information, provide a basis for examining not only

changes in life expectation and health, but also the nature of the changing pattern of diseases over time. These studies had been supplemental by other longitudinal data-sets, including sampling of births and of babies born between 1910 and 1934, to examine inter-generational factors in health and longevity. As a result of these studies Fogel has become involved in the analysis of the economics of aging and longevity, the impact of the expansion of leisure time in the developed world, and the increasing burden of health care and the complexities of achieving equity in health care in recent years, arguing that these issues reflect social and economic progress, not new difficulties. He has also estimated, based on the work of Dora Costa (2003) and others, the magnitude of the continued increase in the length of life in the twentieth century.

Fogel has also made other important contributions to the study of economics and history, to the study of methodology in the social sciences and in history, as seen in his debate with Geoffrey Elton, *Which Road to the Past? Two Views of History* (1983), and to the study of the relations among religious, economic, and political changes. His 2000 book, *The Fourth Great Awakening and the Future of Egalitarianism*, studied the emergence of a religious belief in egalitarianism over time, and how the periodic bursts of awakening influenced the political and economic worlds, as well as what were the major changes in measured inequality in the United States over time.

In addition to his numerous contributions to economics and economic history, Fogel has been a leading figure in proselytizing for cliometrics in economics and in history (including the publication of a 1971 collection of cliometric essays, *The Reinterpretation of American Economic History*, coedited with Stanley Engerman), has been influential in advocating large-scale data collection and analysis, and has been a major producer of scholars for the next generation of economic history. Honours, besides the Nobel Prize, include membership in the National Academy of Science and the American Academy of Arts and Sciences; presidencies of the Economic History Association, the Social Science History Association, and the American Economic Association; the Bancroft Prize; and the Pitt Professorship of American History and Institutions at the University of Cambridge. He was the first director of the Development of the American Economy Program of the National Bureau of Economic Research, chairman of the Committee on Mathematical and Statistical Methods in History of the Mathematical Social Science Board, and is presently the Director of the Center for Population Economics at the University of Chicago.

F

See Also

- ▶ Anthropometric History
- ▶ Cliometrics
- ▶ Economic History
- ▶ Population Ageing

Selected Works

1960. *The Union Pacific railroad: A case in premature enterprise*. Baltimore: Johns Hopkins Press.

1964. *Railroads and American economic growth: Essays in econometric history*. Baltimore: Johns Hopkins Press.

1971. (With S. Engerman, eds.) *The reinterpretation of American economic history*. New York: Harper and Row.

1974. (With S. Engerman.) *Time on the cross: The economics of American Negro Slavery*. Boston: Little, Brown.

1974. (With S. Engerman.) *Time on the cross: Evidence and method, a supplement*. Boston: Little, Brown.

1983. (With G. Elton.) *Which road to the past? Two views of history*. New Haven: Yale University Press.

1989. *Without consent or contract: The rise and fall of American Slavery*. New York: Norton.

2000. *The fourth great awakening and the future of egalitarianism*. Chicago: Chicago University Press.

2004. *The escape from hunger and premature death, 1700–2100: Europe, America, and the third world*. Cambridge: Cambridge University Press.

Bibliography

Conrad, A., and J. Meyer. 1958. The economics of slavery in the Antebellum South. *Journal of Political Economy* 66: 95–130.

Costa, D. 2003. *Health and labor force participation over the life cycle: Evidence from the past*. Chicago: University of Chicago Press.

Fishlow, A. 1965. *American railroads and the transformation of the Ante-Bellum economy*. Cambridge, MA: Harvard University Press.

Floud, R., K. Wachter, and G. Annabel. 1990. *Height, health, and history: Nutritional status in the United Kingdom, 1750–1980*. Cambridge: Cambridge University Press.

Gallman, R. 1970. Self-sufficiency in the cotton economy of the Antebellum South. In *The structure of the cotton economy of the Antebellum South*, ed. W. Parker. Washington, DC: Agricultural History Society.

Goldin, C., and H. Rockoff (eds.). 1992. *Strategic factors in nineteenth century American history: A volume to honor Robert W. Fogel*. Chicago: Chicago University Press.

Komlos, J. (ed.). 1994. *Stature, living standards, and economic development: Essays in Anthropometric history*. Chicago: University of Chicago Press.

Parker, W. (ed.). 1970. *The structure of the cotton economy of the Antebellum South*. Washington, DC: Agricultural History Society.

Steckel, R. 1995. Stature and the standard of living. *Journal of Economic Literature* 33: 1903–40.

Steckel, R., and R. Floud (eds.). 1997. *Health and welfare during industrialization*. Chicago: University of Chicago Press.

Forbonnais, François Véron Duverger de (1722–1800)

Peter Groenewegen

Keywords

Bimetallism; Forbonnais, F. V. G. de; Gournay, Marquis de; Mathematics and economics; Physiocracy; Quesnay, F.

JEL Classifications

B31

French economist, industrialist and inspector of commerce, Forbonnais was born at Le Mans in 1722 and died in Paris in 1800. After initial employment in industry and trade in Nantes, his desire to obtain an official position in the government services (successful in 1756 when he was appointed general inspector of currency) inspired his career as a writer on economic and financial subjects. These all have a strong mercantilist flavour, and also display considerable antagonism to the Physiocrats. Forbonnais contributed a number of economic articles to the *Encyclopédie* and provided translations of some important writings on commerce. These include King's *The British Merchant* (1721) and Uztariz's *Theory and Practice of Commerce* (1724), the former translation according to Morellet (1821) inspired by Gournay.

Forbonnais' major works are his *Elémens du commerce* (1754) and his *Principes et observations oeconomiques* (1767). The *Elémens* has the distinction of being the first French work on economics using mathematical argument. This is his analysis of equilibrium conditions with respect to the rates of exchange between more than two countries and in situations of bimetallism where there are differences in the price ratios of gold and silver (Theocharis 1961). The *Principes* is a polemical work in which the major part is devoted to criticism of Quesnay's *Tableau économique* and his *Encyclopédie* articles on Farmers and Corn after an elucidation of general principles. Forbonnais' criticism of Physiocratic analysis is noteworthy because it was directed at its empirical foundations. In the discussion of general principles he develops arguments on the interdependence of production and trade, the balance of trade, the balance of trade doctrine in relation to money supply and employment, the beneficial consequences of gradual price rises, and the advantages of paper credit.

Selected Works

1754. *Eléments du commerce*. Leyden/Paris: Briasson.

1767. *Principes et observations oeconomiques*. Amsterdam/Paris: M.M. Rey.

Bibliography

King, C. 1721. *The British merchant.* Translated freely from the English by F.V.D. de Forbonnais as *Le négociant anglais,* Dresden/Paris, 1753.
Morellet, l' Abbé de. 1821. *Mémoires de l'Abbé Morellet.* Paris: Librairie Française.
Theocharis, R.D. 1961. *Early developments in mathematical economics*. London: Macmillan.
Uztariz, G. de. 1724. *Theory and practice of commerce* [*Téorica y práctica del commercio*]. Translated freely from the Spanish by F.V.D. de Forbonnais as *Théorie et pratique du commerce et de la marine* Paris, 1753.

Forced Saving

Björn Hansson

Abstract

The forced saving doctrine proposes that an increase in the amount of money may be favourable to capital accumulation at the cost of a reduction in consumption of certain individuals, who have not saved voluntarily. A consensus emerged that new credit might lead to additional, at least temporary, investment even in a full employment situation via an increase in the price level, though Lindahl and Keynes did not consider the extra saving to be forced. However, it was generally thought unwise and unjust to rely on credit inflation as a means of increasing capital accumulation.

Keywords

Bentham, J.; Bullionist Controversy; Business cycle; Capital accumulation; Credit cycle; Forced saving; Hayek, F. A. von; Inflation; Keynes, J. M.; Lindahl, E. R.; Malthus, T. R.; Mises, L. E. von; Over-investment; Ricardo, D.; Robertson, D. H.; Saving equals investment; Thornton, H.; Voluntary saving; Wicksell, J. G. K

JEL Classifications

B1; B2

The doctrine of forced saving proposes that an increase in the amount of money may be favourable to capital accumulation at the cost of a reduction in consumption of certain individuals, but the latter have not saved voluntarily and they do not receive any immediate benefit. The doctrine was developed in the early 19th century by Thornton (1802) and Bentham (1804). They used the terms 'defalcation of revenue' and 'forced frugality' respectively. It was Mises who coined the term 'forced saving' (*erzwungenes Sparen*).

Thornton published his *Paper Credit* (1802) during the debate on the suspension of gold payments by the Bank of England in 1797; the debate concerned the possible existence of a natural tendency to keep the circulation of the Bank of England within the limits which would prevent a dangerous depreciation. An excessive issue of paper money could, according to Thornton, at least temporarily increase the price level of commodities while the money wage and other fixed incomes stayed the same. This would not only lead to a general rise in prices but also to some increase in real capital, since the real consumption of the labourers and recipients of fixed incomes would be reduced, which was the meaning of 'defalcation of revenue'.

Jeremy Bentham, in the manuscript 'Institute of Political Economy' of 1804, some of which had already been written in the years 1800 and 1801, analysed the effects of an increase of paper money in a situation where all hands were employed in the most advantageous manner. If the money in the first instance were used for productive expenditure, that is, buying inputs for producing capital goods, then it would add to real capital. In the second round the money would be exclusively used for consumption and only prices would be affected. The extra real capital was due to the 'forced frugality' of the possessors of fixed income which was engineered by the decrease in the value of money; it operated exactly like an indirect tax upon pecuniary income. But the effect of 'forced frugality' was probably quite small. It was also an unjust mechanism for increasing national wealth, and under normal circumstances voluntary sacrifices would be sufficient to

augment the mass of real wealth. It is obvious in these early enquiries that the forced saving by receivers of fixed incomes came from a decrease in the amount of their real consumption, while the total amount of their money expenditures was kept the same and there was no change in the amount of hoarded funds.

During the course of the Bullionist Controversy, Malthus raised the issue in his 1811 review of Ricardo's *High Price of Bullion* (1810). Malthus proposed that if a new issue of notes came into the hands of the productive classes (described as a change in the distribution of the circulating medium), then capital accumulation would increase. The mechanism of forced saving worked via the increase in the price level, which reduced the share of the annual produce of those classes who were only buyers and not sellers. Ricardo replied, in an appendix to the fourth edition of *The High Price of Bullion* published in 1811, that Malthus's results were based upon the assumption that those who lived on fixed incomes must consume their whole income. In the case of money saving it was possible that the issue of banknotes and the ensuing inflation merely transferred saving from the receivers of fixed incomes to those who had borrowed from the banks. Thus Ricardo saw no reason why it should add anything to the productive classes.

Later, comments on forced saving are found in the works of J.S. Mill and Walras, but the doctrine became important once again when it was incorporated into the pre-Keynesian analysis of credit and business cycles. The analysis took off from Wicksell's brief mention that during a cumulative process rising prices might force people living on fixed money income to reduce their consumption, an 'involuntary saving' which could lead to the production of new real capital. Mises (1912) and later Hayek (1929) developed Wicksell's analysis, and forced saving was used to explain the upswing in the so-called 'over-investment' theories of cyclical movements. An overextension of credit, since the money rate of interest was too low, and the ensuing cumulative process led to a distortion of the vertical structure of production. Production of producers' goods outstripped the production of consumers' goods since means of production were transferred from the latter to the former. The increase in real capital took place because of forced saving, which worked through prices rising faster than disposable income of wage-earners and the rigidity of certain incomes. The intermediate result was the same as for voluntary saving. Consumers were forced to forgo what they used to consume so as to give the entrepreneurs, who had received the additional money, command over resources for the production of extra capital goods. However, no permanent increase of real capital was possible with the help of inflationary credit expansion and forced saving, and the new capital built during the upswing would necessarily be destroyed during the downturn.

Dennis Robertson made a most detailed analysis of different forms of saving or 'lacking' in *Banking Policy and the Price Level* (1926). He introduced the term 'automatic lacking': an involuntary reduction in planned consumption, which came about when the price level increased because newly created money was added to the daily stream of money which competed for the daily stream of marketable goods.

Parts of the doctrine of forced saving were questioned with the publication of Keynes's *Treatise on Money* (1930) and his subsequent debate with Hayek and Robertson. Robertson had, according to Keynes, no distinct definition of voluntary saving, which was related to a confusion concerning the definition of income, and it implied a deficient view of the meaning of forced saving. Keynes defined saving as the difference between income or normal costs and expenditure on consumption, which could differ from investment since saving and investment were decisions taken by different agents, windfall profits and losses being the balancing figure between investment and saving. Forced saving or automatic lacking existed when investment exceeded saving and purchasing power was redistributed by the accompanying inflation; it was represented on the one hand by the increased amount of money which spenders had to pay for that part of consumption which they continued to enjoy, and on the other hand by the extra investment provided out of the windfall gains of the entrepreneurs.

Hence Keynes did not challenge the fact that an increase in net investment took place via the redistribution of purchasing power, but it was not an involuntary act.

At the same time, Erik Lindahl presented a similar analysis in *The Rate of Interest and the Price Level* (1930). The rising prices during an upward cumulative process had to change the distribution in favour of those who had a strong incentive to save, until the total saving in the community corresponded to the value of real investment, which was primarily determined by the rate of interest. This saving was mainly voluntary, since an individual was free to consume as much as he liked and the only limit was his credit standing. Keynes had the same view in the *General Theory*: this type of saving was in complete agreement with the free will of the individual to save what he chose irrespective of what he or others might be investing, since no individual could be compelled to own the additional money (corresponding to the new bank-credit) unless he deliberately preferred to hold more money rather than some other form of wealth. Lindahl reserved forced saving for the possibility that the individual has to limit planned consumption out of income (defined as the rate of interest on the capital value of all capital goods including human capital) because he is not able to obtain credit, which might be explained by banking rules concerning the collateral for loans, i.e. it is not a perfect capital market.

Once the notions of *ex ante* and *ex post* were introduced all these problems could be solved. A fall in the money rate leads to an excess of planned and realized investment over planned saving (related to planned income), and the subsequent increase in prices would imply higher incomes *ex post* for the entrepreneurs, which is the same as Keynes's concept of windfall gains in the *Treatise.* This unexpected windfall, which could not be spent during the period, would contribute the extra necessary saving, since investment *ex post* had to be equal to saving *ex post*. Lindahl denoted this as 'unintentional saving' and he found 'forced saving' to be an inappropriate term. However, Keynes seemed to have changed his position slightly in *How to Pay for the War* (1940). The process could be successful only if wages lagged behind prices, for otherwise an unlimited inflation would take place. As such it was a method of compulsorily converting a part of workers' earnings, which they do not plan to save voluntarily, into the voluntary saving of the entrepreneurs. From an analytical point it was voluntary saving, but it was 'a matter of taste' whether this was a suitable name.

To sum up: there was a consensus that new credit might lead to an additional, at least temporary, investment even in a full employment situation via an increase in the price level. But the most recent contributions – for example, Lindahl and Keynes – did not consider the extra saving to be forced. At the same time almost all of them found it unwise and unjust to rely on credit inflation as a means of increasing capital accumulation. However, after Keynes's analysis in the *General Theory* the problem seems to have disappeared from the agenda.

See Also

► Inflation

Bibliography

Bentham, J. 1804. Institute of political economy. In *Jeremy Benthams economic writings,* vol. 3, ed. W. Stark. London: George Allen & Unwin, 1954.

Haberler, G. 1937. *Prosperity and depression,* 5th ed. London: George Allen & Unwin, 1964.

von Hayek, F. 1929. *Geldtheorie und Konjunkturtheorie.* Vienna/Leipzig: Holder-Pichler-Tempsky. Trans. as *Monetary theory and the trade cycle,* London: Jonathan Cape, 1933.

von Hayek, F. 1931. *Prices and production,* 2 ed. London: George Routledge & Sons, 1935.

von Hayek, F. 1932. A note on the development of the doctrine of forced saving. *Quarterly Journal of Economics* 47 (November): 123–133.

Keynes, J.M. 1930. A treatise on money, vol. 1. In *The collected writings of John Maynard Keynes*, vol. 5. London: Macmillan, 1971.

Keynes, J.M. 1936. The general theory of employment, interest and money. *In The collected writings of John Maynard Keynes,* vol. 8. London: Macmillan, 1973.

Keynes, J.M. 1940. How to pay for the war. *In the collected writings of John Maynard Keynes,* vol. 22. London: Macmillan, 1978.

Lindahl, E. 1930. *Penningpolitikens medel.* Lund: C.W.K. Gleerup. Trans. as "The rate of interest and the price level" in E. Lindahl, *Studies in the theory of money and capital.* London: George Allen & Unwin, 1939.

Machlup, F. 1943. Forced or induced saving: An exploration into its synonyms and homonyms. *The Review of Economics and Statistics* 25: 26–39.

Malthus, T.R. 1811. Review of Ricardo's *high price of bullion. Edinburgh Review,* February.

Mill, J.S. 1844. Essays on some unsettled questions of political economy. *In collected works of John Stuart Mill,* vol. 4. London: Routledge & Kegan Paul, 1967.

von Mises, L. 1912. *Theorie des Geldes undder Umlaufsmittel.* Munich: Dunker & Humblot. 2nd edn, 1924. Trans. as *The theory of money and credit.* London: Jonathan Cape, 1934.

Ricardo, D. 1810. The high price of bullion. A proof of the depreciation of bank notes. In *The works and correspondence of David Ricardo,* vol. 3, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.

Robertson, D. 1926. *Banking policy and the price level*. London: P.S. King & Sons.

Thornton, H. 1802. *An enquiry into the nature and effects of the paper credit of great britain.* Reprinted. London: George Allen & Unwin, 1939.

Walras, L. 1879. Théorie mathématique du billet de banque. Reprinted in *Etudes d'économie politique appliquée,* Lausanne/Paris, 1898.

Wicksell, K. 1935. *Lectures on political economy,* vol. 2. New York: Kelley. Trans. from the 3rd Swedish edn of *Forelasningar i nationalekonomi,* vol. 2. London: George Routledge & Sons, 1929.

Forecasting

Clive Granger

Abstract

Providing timely and useful forecasts is among the most relevant tasks of economists. The choice among the many techniques and approaches depends on the variables being forecast and the length of the forecast horizon. Providing confidence intervals around the point forecasts is becoming standard practice, as are sophisticated attempts at evaluating the quality of the forecasts and the intervals.

Forecasts are often combined, raising questions about the appropriate cost functions to use in the evaluation process. Economists once concentrated on forecasting the mean of a process, then moved to variance, and now consider quantities and the whole distribution.

Keywords

Akaike information criterion; ARCH models; ARMA models; Bayes information criterion; Copulas; Error-correction models; Forecasting; Kalman filters; Leading indicators; Linear models; Neural networks; Quantiles; Switching models; Time series analysis; Vector autoregressions

JEL Classifications

C53

Decisions in the fields of economics and management have to be made in the context of forecasts about the future state of the economy or market. As decisions are so important as a basis for these fields, a great deal of attention has been paid to the question of how best to forecast variables and occurrences of interest. There are several distinct types of forecasting situations, including event timing, event outcome, and time-series forecasts. Event timing is concerned with the question of when, if ever, some specific event will occur, such as the introduction of a new tax law, or of a new product by a competitor, or of a turning point in the business cycle. Forecasting of such events is usually attempted by the use of leading indicators, that is, other events that generally precede the one of interest. Event outcome forecasts try to forecast the outcome of some uncertain event that is fairly sure to occur, such as finding the winner of an election or the level of success of a planned marketing campaign. Forecasts are usually based on data specifically gathered for this purpose, such as a poll of likely voters or of potential consumers. There clearly should be a positive relationship between the amount spent on gathering the extra data and the quality of the forecast achieved.

A time series x_t is a sequence of values gathered at regular intervals of time, such as daily stock market closing prices, interest rates observed

weekly, or monthly unemployment levels. Irregularly recorded data, or continuous time sequences may also be considered but are of less practical importance. When at time n (now), a future value of the series, x_{n+h}, is a random variable where h is the forecast horizon. It is usual to ask questions about the conditional distribution of x_{n+h} given some information set I_n, available now from which forecasts will be constructed. Of particular importance are the conditional mean

$$f_{n,h} = E[x_{n+h}|I_n]$$

and variance, $V_{n,h}$. The value of $f_{n,h}$ is a point forecast and represents essentially the best forecast of the most likely value to be taken by the variable x at time $n + h$.

With a normality assumption, the conditional mean and variance can be used together to determine an interval forecast, such as an interval within which $x_{n,h}$ is expected to fall with 95 per cent confidence. An important decision in any forecasting exercise is the choice of the information set I_n. It is generally recommended that I_n include at least the past and present of the individual series being forecast, $x_{n-j,j\geq 0}$. Such information sets are called *proper*, and any forecasting models based upon them can be evaluated over the past. An I_n that consists just of x_{n-j}, provides a univariate set so that future x_i are forecast just from its own past. Many simple time-series forecasting methods are based on this information set and have proved to be successful. If I_n includes several explanatory variables, one has a multivariate set. The choice of how much past data to use and which explanatory variables to include is partially a personal one, depending on one's knowledge of the series being forecast, one's levels of belief about the correctness of any economic theory that is available, and on data availability. In general terms, the more useful are the explanatory variables that are included in I_n, the better the forecast that will result. However, having many series allows for a confusing number of alternative model specifications that are possible so that using too much data could quickly lead to diminishing marginal returns in terms of forecast quality. In practice, the data to be used in I_n will often be partly determined by the length of the forecast horizon. If h is small, a short-run forecast is being made and this may concentrate on frequently varying explanatory variables. Short-term forecasts of savings may be based on interest rates, for example. If h is large so that long-run forecasts are required, then slowly changing, trending explanatory variables may be of particular relevance. A long-run forecast of electricity demand might be largely based on population trends, for example. What is considered short run or long run will usually depend on the properties of the series being forecast. For very long forecasts, allowances would have to be made for technological change as well as changes in demographics and the economy. A survey of the special and separate field of technological forecasting can be found in Martino (1993) with further discussion in Martino (2003).

If decisions are based on forecasts, it follows that an imperfect forecast will result in a cost to the decision-maker. For example, if $f_{n,h}$ is a point forecast made at time n, of x_{n+h}, the eventual forecast error will be

$$e_{n,h} = x_{n,h} - f_{n,h},$$

which is observed at time $n + h$. The cost of making an error e might be denoted as $C(e)$, where $C(e)$ is positive with $C(0) = 0$. As there appears to be little prospect of making error-free forecasts in economics, positive costs must be expected, and the quality of a forecast procedure can be measured as the expected or average cost resulting from its use. Several alternative forecasting procedures can be compared by their expected costs and the best one chosen. It is also possible to compare classes of forecasting models, such as all linear models based on a specific, finite information set, and to select the optimum model by minimizing the expected cost. In practice the true form of the cost function is not known for decision sequences, and in the univariate forecasting case a pragmatically useful substitute to the real $C(e)$ is to assume that it is well approximated by ae^2 for some positive a. This enables least-squares statistical techniques to be used when a model is estimated and is the basis of a number of

theoretical results including that the optimal forecast of x_{n+h} based on I_n is just the conditional mean of $x_{n,h}$. Machina and Granger (2006) have considered cost functions generated by decision makers and then find implications for their utility functions. This is just one component of considerable developments in the area of evaluation of forecasts; see West (2006) and Timmermann (2006), for example.

When using linear models and a least-square criterion, it is easy to form forecasts under an assumption that the model being used is a plausible generating mechanism for the series of interest. Suppose that a simple model of the form

$$x_t = \alpha x_{t-1} + \beta y_{t-2} + \varepsilon_t$$

is believed to be adequate where ε_t is a zero-mean, white noise (unforecastable) series. When at time n, according to this model, the next value of x will be generated by

$$x_{n+1} = \alpha x_n + \beta y_{n-1} + \varepsilon_{n+1}.$$

The first two terms are known at time n, and the last term is unforecastable. Thus

$$f_{n,1} = \alpha x_n + \beta y_{n-1}$$

and

$$e_{n,1} = \varepsilon_{n+1}.$$

x_{n+2}, the following x, will be generated by

$$x_{n+2} = \alpha x_{n+1} + \beta y_n + \varepsilon_{n+2}.$$

The first of these terms is not known at time n, but a forecast is available for it, αf_n; the second term is known at time n, and the third term is not forecastable, so that

$$f_{n,2} = \alpha f_{n,1} + \beta y_n$$

and

$$e_{n,2} = \varepsilon_{n+2} + \alpha\left(x_{n+1} - f_{n,1}\right) = \varepsilon_{n+2} + \alpha\varepsilon_{n+1}.$$

To continue this process for longer forecast horizons, it is clear that forecasts will be required for y_{n+h-2}. The forecast formation rule is that one uses the model available as though it is true, asks how a future x_{n+h} will be generated, uses all known terms as they occur, and replaces all other terms by optimal forecasts. For non-linear models this rule can still be used, but with the additional complication that the optimum forecast of a function of x is not the same function of the optimum forecast of x.

The steps involved in forming a forecast include deciding exactly what is to be forecast, the forecast horizon, the data that is available for use, the model forms or techniques to be considered, the cost function to be used in the evaluation procedure, and whether just one single forecast would be produced or several alternatives. It is good practice to decide on the evaluation to be used before starting a sequence of forecasts. If there are several alternative forecasting methods involved, a weighted combination of the available forecasts is both helpful for evaluation and can often provide a superior forecast.

The central problem in practical forecasting is choosing the model from which the forecasts will be derived. If a univariate information set is used, it is natural to consider the model developed in the field of time-series analysis. A class of models that has proved to be successful in short-term forecasting is the autoregressive (AR) model class. If a series is regressed on itself up to p lags, the result is an AR (p) model. These models were originally influenced by Box and Jenkins (1970) as a particularly relevant subclass of their ARMA (p, q) models, which involve moving average components. The number of lags in an AR(p) can be chosen using a selection criterion; the most used are the Bayes information criterion (BIC) and the less conservative Akaike information criterion (AIC).

The natural extension was to vector autoregressive models. Later, when it was realized that many series in macroeconomics and finance had the property of being integrated, and so contained stochastic trends, the natural multivariate form was the error-correction model. It is quite often found that error-correction models improve forecasts, but not inevitably. There are a variety of ways of building

models with many predictive variables, including those with unobserved components and using special data, such as survey expectations, real-time macro data, and seasonal components.

In recent years the linear models have been joined by a variety of nonlinear forms (see Terasvita 2006), including switching models and neural networks as well as linear models with time varying coefficients estimated using Kalman filters.

Traditionally, forecasters concentrated on the mean of the predictive distribution. Towards the end of the 20th century considerable attention was given to forecasting the variance of the distribution, particularly in the financial area, often using Engle's (1995) ARCH model or one of its many generalizations (see the survey by Andersen et al. 2006). Recently forecasts of the whole distribution have become more common in practice, both in finance and in macroeconomics: see Corradi and Swanson (2006) for a recent discussion. These forecasts will include discussions of quantiles, and the use of copulas gives a way into multivariate distribution forecasts. The topics mentioned in this paragraph are covered by chapters in Elliott et al. (2006).

Bibliography

Andersen, T., T. Bollerslev, P. Christoffersen, and F. Diebold. 2006. Volatility and correlation forecasting. In Elliott, Granger and Timmermann (2006).

Box, G., and G. Jenkins. 1970. *Time series analysis, forecasting and control*. San Francisco: Holden Day.

Clements, M., and D. Hendry. 2003. *Forecasting economic time series*. Cambridge: Cambridge University Press.

Corradi, V., and N.R. Swanson. 2006. Predictive density evaluation. In Elliott, Granger and Timmermann (2006).

Elliott, G., C.W.J. Granger, and A. Timmermann (eds.). 2006. *Handbook of economic forecasting*. Amsterdam: North-Holland.

Engle, R.F. 1995. *ARCH: Selected readings*. Oxford: Oxford University Press.

Granger, C.W.J., and P. Newbold. 1987. *Forecasting economic time series*, 2nd ed. New York: Academic.

Machina, M., and C.W.J. Granger. 2006. Forecasting and decision theory. In Elliott, Granger and Timmermann (2006).

Martino, J.P. 1993. *Technological forecasting for decision making*, 2nd ed. New York: McGraw Hill.

Martino, J.P. 2003. A review of selected recent advances in technological forecasting. *Technological Forecasting and Social Change* 70: 719–733.

Terasvita, T. 2006. Forecasting economic variables with nonlinear models. In Elliott, Granger and Timmermann (2006).

Timmermann, A. 2006. Forecast combinations. In Elliott, Granger and Timmermann (2006).

West, K. 2006. Forecast evaluation. In Elliott, Granger and Timmermann (2006).

Foreclosure, Economics of

F

K. Gerardi

Abstract

This entry describes the economics of foreclosure with respect to US residential mortgage markets from the perspective of both the borrower and the lender.

Keywords

Deed-in-lieu; Default; Equity; Forbearance; Foreclosure; Loan modification; Mortgage; Short-sale

JEL Classifications

D11; D12; G21

Foreclosure

Foreclosure is the legal process by which a lender repossesses a home from a borrower. Legally, a mortgage is a type of repurchase agreement which transfers ownership of the property from the borrower to the lender, but gives the borrower the right to buy the property back by paying the outstanding balance on the mortgage. In the event that the borrower defaults on her obligations to the lender by missing periodic loan payments, the lender can extinguish or *foreclose* on the borrower's right to repurchase the property.

This description is oversimplified, and the precise legal status of the lender's ownership stake depends on the type of mortgage and the jurisdiction, but the principle is always the same.

The foreclosure process starts when the borrower defaults on the promissory note, typically by missing a payment, although any violation of the contract – renting the property, for example – may constitute a default. The lender then has the right to demand full repayment or, in legal jargon, to accelerate the mortgage. Common law generally allows the borrower a period to correct the default and resume making periodic payments. Typically, this breathing space, known as the period of equitable redemption, lasts three months, after which, the lender has the right to foreclose. Even after the equitable redemption period, and in some cases after the legal foreclosure, the borrower still has the right to redeem the mortgage by repaying the loan in full including all arrears, fees, taxes and penalties.

In the USA there are two varieties of foreclosure: judicial and non-judicial. Some states allow both types of foreclosure and some allow only one or the other. Under judicial foreclosure, the lender must file a suit to initiate the foreclosure process. Under non-judicial foreclosure, the lender initiates the foreclosure process by exercising a power of sale clause without having to go to court. In most cases the lender will try to sell the property at public auction and use the proceeds to pay off the outstanding mortgage debt and any fees incurred from the foreclosure process. If the highest bid at the auction does not meet the lender's reservation price, then the lender will legally repossess the property. The lender then adds the property to its balance sheet and puts the property up for sale through normal channels.

Foreclosure is not the only remedy the lender has to recover the obligations of the borrower in the promissory note. If the proceeds from the sale of the property fall short of those obligations, lenders can seek to recover the difference. Outside the USA, lenders generally have substantial powers to do this, while in the USA, the ability of lenders to obtain deficiency judgments (unsecured claims for the gap) depends on the state.

The Borrower's Decision to Default

Economists generally model default as an option embedded in a mortgage contract. In the simplest theoretical setting, default gives the borrower the option to sell the house back to the lender for the outstanding balance of the mortgage. The borrower exercises this option by stopping payment on the mortgage. The academic literature on mortgage default has largely considered the borrower's default decision to be similar to an investor's decision on whether or not to exercise a financial option. Many studies, beginning in the 1980s, such as Cunningham and Hendershott (1984) and Epperson et al. (1985) used the option-based valuation models pioneered by Black and Scholes (1973) to study the default decision.

The default option model has been the source of some confusion among researchers and policy makers. What the model says is that the borrower should exercise the option when the *value* of the mortgage exceeds the *value* of the house. But many assume that the value of the mortgage equals the unpaid principal balance and then interpret the model as implying that any borrower with an unpaid principal balance that exceeds the value of the house, that is, who has negative equity, should default. But this interpretation is incorrect, since it ignores the value to the borrower of exercising *future* default and prepayment (repurchase) options, which are forfeited once the borrower defaults. The options to default or prepay in the future reduce the true cost of the mortgage to a level that is below the remaining principal balance. Consequently, a borrower with negative equity may benefit from waiting to exercise the default option.

The first generation of option-based valuation models assumed that all borrowers were identical (based on an assumption of perfect capital markets), and attempted to estimate the equity threshold at which default would occur (Kau et al. 1994). But the assumption of an identical threshold across borrowers is contradicted in the data. As a result, the literature has stressed the idea that there is something unaccounted for by these models that creates a significant amount of heterogeneity across borrowers in their decision to exercise the default option. An explanation involving heterogeneous

transaction costs to defaulting emerged in the literature. Transaction costs include such factors as future limitations on credit availability, purchasing or sale costs, tax treatment, or even psychological costs to defaulting.

An alternative to the transaction costs explanation of mortgage default, first discussed by Riddiough (1991), posits that 'trigger events' – divorce, illness and spells of unemployment are the typical examples – make some borrowers more vulnerable to default. Gerardi et al. (2007) develop a simple model to formally explain the channel by which trigger events may lead to default. The authors argue that depending on their income prospects, financial situation and other factors, borrowers discount the future differently. The cost of funds is the relevant rate at which borrowers discount future payoffs and consumption, since it is the rate at which a borrower is willing to sacrifice future consumption for current consumption. The relevant cost of funds for a borrower with credit card debt for example is the credit card interest rate, while the cost of funds for a borrower with only riskless savings is the return on riskless savings. Differences in the cost of funds across borrowers are correlated with the individual-level shocks discussed above, because borrowers in financial distress are much more likely to borrow at high interest rates, and thus discount future consumption to a greater extent than financially sound borrowers. Since financially stressed borrowers discount future consumption at a high rate, they are more likely to default in order to increase current consumption (by the amount of the mortgage payment). Thus, the cost of funds provides a channel for the link between employment shocks, medical shocks and even family level shocks such as divorce, and the incidence of default.

The Lender's Decision to Foreclose

When a borrower defaults, foreclosure is only one of many options that a mortgage lender can pursue. The foreclosure process typically imposes very high costs on the lender, including the opportunity cost of principal and income not received; additional servicing, legal and property maintenance expenses; and costs associated with property disposition, which often increase substantially during housing market downturns as demand shrinks and houses become harder to sell. As a result of these costs, lenders often have an incentive to explore alternatives to foreclosure.

An alternative to foreclosure that received a great deal of attention during the housing crisis of the mid-to-late 2000s is loan modification. A loan modification occurs when the lender permanently changes at least one of the terms of the mortgage contract (such as the interest rate, maturity date or remaining principal balance), usually in the favour of the borrower, so as to increase the probability that the borrower repays the mortgage. Another alternative to foreclosure is a preforeclosure, or 'short' sale, in which the lender allows the borrower to sell the house to a third party at a price below the outstanding mortgage balance (inclusive of sale costs and other fees). The lender can then negotiate an unsecured repayment plan with the borrower for the additional amount owed or can forgive the remaining debt outright (Cutts and Green 2004). Another foreclosure alternative, called a 'deed-inlieu', occurs when the borrower voluntarily surrenders the title of the house back to the lender in exchange for a release from all mortgage obligations. Relative to foreclosure, deeds-in-lieu reduce the time in which a borrower who has defaulted can live 'rent free' in the house relative to foreclosure, but they are often less costly to the borrower in terms of reduced access to future credit.

A foreclosure alternative that often works well for borrowers undergoing temporary liquidity problems is forbearance. In this case, the lender agrees not to foreclose for some given time period, during which the lender receives reduced payments from the borrower. The forbearance period is designed to be long enough to allow the borrower to find a new job or otherwise correct his or her financial problems. In return, the borrower agrees to a mortgage repayment plan that will, over a specific time period, bring the borrower current on the mortgage again. Springer and Waller (1993) explore the use of forbearance as a loss mitigation tool, while Foote et al. (2008) discuss the benefits of forbearance over loan modification when the potential default is caused by trigger events.

See Also

► Rent

Bibliography

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81(3): 637–654.

Cunningham, D., and P.H. Hendershott. 1984. Pricing FHA mortgage default insurance. *Housing Finance Review* 3(4): 373–392.

Cutts, A.C. and Green, R.K. 2004. Innovative servicing technology: Smart enough to keep people in their houses? *Freddie Mac Working Paper* 04–03.

Epperson, J., J.B. Kau, D.C. Keenan, and W.J.I.I.I. Muller. 1985. Pricing default risk in mortgages. *AREUEA Journal* 13(3): 152–167.

Foote, C., K. Gerardi, and P. Willen. 2008. Negative equity and foreclosure: Theory and evidence. *Journal of Urban Economics* 64(2): 234–245.

Gerardi, K., A. Shapiro, and P. Willen. 2007. Subprime outcomes: Risky mortgages, homeownership experiences, and foreclosures. *Federal Reserve Bank of Boston Working Paper* 07–15.

Kau, J.B., D.C. Keenan, and T. Kim. 1994. Default probabilities for mortgages. *Journal of Urban Economics* 35(3): 278–296.

Riddiough, T. 1991. Equilibrium mortgage default pricing with non-optimal borrower behavior, *PhD Dissertation*, University of Wisconsin Madison.

Springer, T.M., and N.G. Waller. 1993. Lender forbearance: Evidence from mortgage delinquency patterns. *AREUEA Journal* 21(1): 27–46.

Foreign Aid

Tarp Finn and Hollis Chenery

Abstract

Foreign aid has evolved significantly since the Second World War in response to a dramatically changing global political and economic context. This article (a) reviews this process and associated trends in the volume and distribution of foreign aid; (b) reviews the goals, principles and institutions of the aid system; and (c) discusses whether aid has been effective. While much of the original optimism about the impact of foreign aid needed modification, there is solid evidence that aid has indeed helped further growth and poverty reduction.

Keywords

Bretton Woods agreement; Charity; Chenery, H.; Corruption; Dependency; Development Assistance Committee (DAC); Dutch disease; Economic development; Foreign aid; Foreign direct investment; Harrod-Domar model; International Development Association (IDA); International public goods; Krueger, A.; Marshall Plan; Micro-macro paradox; Millennium Challenge Account (MCA); Millennium Development Goals (MDG); Non-governmental organizations (NGOs); Official Development Assistance (ODA); Poverty; Rent-seeking; Washington consensus; World Bank

JEL Classifications

O2

Foreign aid and its usefulness in promoting economic development in developing countries has been a topic of intense controversy ever since Rosenstein-Rodan (1943) advocated aid to eastern and south-eastern Europe. Early optimism and confidence in the impact of foreign aid have been tempered with time, but aid continues to loom large in the public discourse; and aid remains squarely on most policy agendas concerned with poverty and inequality in Africa and elsewhere in the developing world.

What is foreign aid? Loosely, it covers governmental transfers to poor countries that are mainly destined for developmental purposes. For a more precise definition it is useful to turn to the Development Assistance Committee (DAC) of the OECD. DAC is the principal body through which the OECD deals with issues related to cooperation with developing countries, and DAC publishes the most comprehensive data available on foreign aid (OECD 2004). DAC countries also account for almost 95 per cent of all aid flows. In 2002 the total amount of foreign aid disbursed by

donors to developing countries and multilateral organizations reached $61.5 billion (Table 1). Multilateral organizations disbursed some 30 per cent (Table 2), and Table 3 shows that international development assistance is an important resource for many developing countries.

The term 'foreign aid' or 'development assistance' refers to financial flows that qualify as Official Development Assistance (ODA). ODA is defined as grants and loans to aid recipients that are: (a) undertaken by the official sector of the donor country, (b) with promotion of economic development and welfare as the main objective, (c) at concessional financial terms, where the grant element is equal to at least 25 per cent.

Conventionally the market rate of interest used to assess a loan is taken as ten per cent. Thus, while the grant element is nil for a loan carrying an interest rate of ten per cent, it is 100 per cent for a pure grant, and lies between these two limits for a soft loan. In addition to financial flows, technical cooperation costs are included in ODA; but grants, loans and credits for military purposes are excluded, and transfer payments to private individuals are in general not counted. The same goes for private charity, hard loans and foreign direct investment (FDI).

While the OECD operates with a consolidated list of recipient countries to capture all aid flows, this list is divided into two parts. Only aid to 'traditional' developing countries counts as ODA. For these (Part I) countries there is a longstanding United Nations (UN) target that they should receive 0.7 per cent of donors' gross national income (GNI) as aid. Assistance to the 'more advanced' eastern European and developing (Part II) countries is recorded separately as 'official aid' (OA), which is not included as part of ODA.

Foreign Aid, Table 1 Net ODA disbursements by donor, 1960–2002

	2002 prices ($ billion)				Per cent of total			
	ODA per capita (2002 prices, $)				Per cent of donor GNI			
	1960–73	1992	1998	2002	1960–73	1992	1998	2002
United States	14.7	14.1	9.4	13.3	47.1	23.0	18.3	21.6
	74.9	55.3	34.8	46.1	0.4	0.2	0.1	0.1
Japan	2.5	10.5	10.4	9.3	8.0	17.1	20.2	15.1
	24.5	84.4	82.2	72.8	0.2	0.3	0.3	0.2
France	3.9	7.2	5.1	5.5	12.8	11.8	9.9	8.9
	80.6	126.2	87.3	92.3	0.8	0.6	0.4	0.4
Germany	2.8	6.6	4.9	5.3	9.1	10.7	9.5	8.7
	48.0	81.4	59.5	64.5	0.4	0.4	0.3	0.3
United Kingdom	3.2	3.6	3.8	4.9	10.2	5.8	7.4	8.0
	58.0	61.3	64.7	83.5	0.5	0.3	0.3	0.3
DK, NL, NO and SE	1.3	7.1	7.4	8.7	4.2	11.5	14.4	14.1
	44.6	211.9	217.0	248.2	0.3	1.0	0.8	0.9
Other DAC	2.6	10.9	9.4	11.3	8.5	17.8	18.2	18.4
	23.0	57.9	46.0	53.6	0.3	0.4	0.3	0.3
Non-DAC		1.4	1.0	3.2		2.2	2.0	5.2
				67.2		0.3	0.1	0.4
Total	31.0	61.3	51.5	61.5	100	100	100	100
	51.6	76.9	61.7	67.6	0.4	0.3	0.2	0.2
Bilateral ODA	26.5	41.9	34.9	43.5	85.5	68.3	67.8	70.7
Multilateral ODA	4.9	19.1	16.6	18.0	15.6	31.1	32.2	29.3

Notes: Denmark (DK), the Netherlands (NL), Norway (NO) and Sweden (SE) reached the UN ODA target of 0.7% of GNI in respectively 1978, 1975, 1976 and 1975. Luxembourg reached the target in 2000
Source: OECD (2004)

Foreign Aid, Table 2 Multilateral aid disbursements, 1960–2002

	2002 prices, $ billion				Per cent			
	1960–73	1992	1998	2002	1960–73	1992	1998	2002
Multilateral, total *of which*:	2.8	16.3	14.4	17.0	100	100	100	100
United Nations	0.9	5.3	2.6	3.8	31.4	32.6	17.9	22.1
IMF and WB	0.8	5.3	5.0	6.0	30.0	32.7	35.0	35.1
European Commission	0.6	3.8	4.6	5.1	23.1	23.1	32.3	30.3
Regional Development Banks	0.4	1.6	1.9	1.8	15.4	10.0	13.2	10.5
Other multilateral institutions	0.0	0.3	0.2	0.4	0.0	1.6	1.6	2.1

Source: OECD (2004)

Historical Background

Foreign aid emerged out of the disruption that followed the Second World War. The international economic system had collapsed, and war-ravaged Europe faced a critical shortage of capital and an acute need for physical reconstruction. The response was the European Recovery Programme, commonly known as the Marshall Plan. During the peak years the United States devoted some two or three per cent of its national income to helping restore Europe. This objective was achieved on schedule, and fuelled optimistic expectations about the future effectiveness of foreign aid.

After the success of the Marshall Plan, the attention of industrialized nations turned to the developing countries, many of which became independent around 1960. Economic growth in a state-led planning tradition became a key objective during the 1950s and 1960s, and it was widely believed that poverty and inequality would eventually be eliminated through growth and modernization ('trickle down'). A major part of the rapidly increasing bilateral flows during the 1950s came from the United States, but colonial ties remained strong, and developing regions continued to receive bilateral (country-to-country) support from the former colonial powers, notably France and the United Kingdom. Yet the 1960s was also the decade when a range of new bilateral donor agencies was established in, for example, the Nordic countries. They accounted for much of the increase in aid flows in the 1970s.

A transition toward more independent, multilateral relations began to emerge during the 1960s. Hjertholm and White (2000) argue that this created, a constituency for foreign aid, and the non-aligned movement of developing countries gave a focus to this voice, as did the various organs of the UN, which accounted for around one-third of multilateral assistance during 1960–73. The International Bank for Reconstruction and Development (IBRD, or World Bank), established at the Bretton Woods Conference in 1944, is central in multilateral development assistance, especially following the creation of the International Development Association (IDA) in 1960. IDA channels resources to the poorest countries on 'soft' conditions alongside the regional development banks, formed from 1959 to 1966, and the European Commission.

The original Marshall Plan was built around support to finance general categories of imports and strengthen the balance of payments (that is, programme aid), but from the early 1950s project aid became the dominating aid modality. Some donors continued to supply programme aid (including food aid), but aid was increasingly disbursed for the implementation of specific capital investment projects and associated technical assistance to support advances in infrastructure and productive sectors.

The multilateralism of aid became somewhat more pronounced after the mid-1970s, when the UN, the World Bank and other multilateral agencies expanded their activities quite considerably; since then the share of multilateral aid in total aid has remained close to 30 per cent. The 1970s also saw an increased focus on employment, income distribution, and poverty alleviation as essential objectives of development and foreign aid. The effectiveness of trickle-down was widely questioned, and new strategies referred to as

Foreign Aid, Table 3 ODA by recipient, 1960–2002

GNI in 2002 (US $ billion)		GNI per capita (2002, US$)	Total ODA receipts (2002 prices – US$ billion)				In per cent of total flows (ODA + OOF + private)			
			ODA per capita (2002 prices, US$)				In per cent of GNI			
			1960–73	1992	1998	2002	1960–73	1992	1998	2002
Developing countries, total			30.2	58.3	49.3	60.5	74.2	55.3	26.6	88.2
Least developed, countries, total			4.1	16.3	12.2	17.8	88.1	96.8	82.9	116.2
Other low-income countries, total			10.7	10.8	10.2	12.3	89.8	63.7	59.7	86.4
Low-middle-income countries, total			6.7	16.9	13.6	16.1	75.0	69.6	31.4	96.3
China	1251.1	977.1		2.8	2.4	1.5		49.2	31.6	−61.9
				2.4	1.9	1.2		0.7	0.3	0.1
Mexico	636.1	6309.3	0.2	0.3	0.0	0.1	23.1	4.4	0.5	2.3
			3.9	3.2	0.4	1.3	0.2	0.1	0.0	0.0
India	506.2	482.7	4.5	2.3	1.6	1.5	98.5	78.4	57.8	5359.6
			9.0	2.6	1.6	1.4	1.9	1.0	0.4	0.3
Brazil	443.0	2538.9	0.9	−0.3	0.3	0.3	69.8	−14.9	1.4	12.3
			10.5	−2.2	1.9	1.9	0.7	−0.1	0.0	0.1
Indonesia	164.6	777.2	1.3	1.8	1.2	1.3	87.3	33.0	18.8	8185.7
			11.8	10.0	6.1	6.2	4.4	1.6	1.4	0.8
Israel	100.9	15365.4	0.5	2.4	1.2	0.8	62.2	64.0	31.3	139.7
			193.4	475.2	193.9	115.3	2.4	3.2	1.1	0.8
Egypt	90.0	1355.3	0.7	3.7	1.9	1.2	79.6	217.3	47.5	63.2
			24.0	68.4	31.3	18.7	2.8	8.7	2.3	1.4
Malaysia	88.4	3639.1	0.1	0.2	0.2	0.1	59.6	16.7	−25.0	2.5
			14.0	10.1	9.1	3.5	0.7	0.4	0.3	0.1
Philippines	83.1	1039.9	0.4	1.7	0.6	0.6	67.2	115.8	14.8	22.8
			12.9	26.0	8.1	6.9	1.0	3.2	0.9	0.7
Colombia	77.8	1778.5	0.4	0.2	0.2	0.4	73.6	288.3	6.8	−19.0
			18.4	6.0	4.1	10.1	1.4	0.5	0.2	0.6
Pakistan	59.8	412.4	1.8	1.0	1.0	2.1	97.1	58.1	62.4	114.0
			33.5	8.6	7.8	14.8	4.5	2.1	1.7	3.6
Bangladesh	49.7	366.6	0.7	1.8	1.1	0.9	98.8	93.0	87.7	102.0
			10.5	15.5	8.8	6.7	5.2	5.6	2.5	1.8
Nigeria	36.9	278.0	0.3	0.3	0.2	0.3	59.2	250.9	61.1	6.6
			6.3	2.5	1.6	2.4	0.9	0.9	0.7	0.9
Guatemala	23.0	1915.9	0.1	0.2	0.2	0.2	99.8	155.8	30.0	94.8
			14.5	22.8	20.9	20.7	1.1	1.9	1.2	1.1

(*continued*)

Foreign Aid, Table 3 (continued)

GNI in 2002 (US $ billion)		GNI per capita (2002, US$)	Total ODA receipts (2002 prices – US$ billion)				In per cent of total flows (ODA + OOF + private)			
			ODA per capita (2002 prices, US$)				In per cent of GNI			
			1960–73	1992	1998	2002	1960–73	1992	1998	2002
Sri Lanka	16.3	858.4	0.2	0.6	0.4	0.3	86.5	91.5	81.3	83.0
			13.1	37.5	23.2	18.1	1.6	6.7	2.7	2.1
Kenya	12.2	389.5	0.3	0.8	0.4	0.4	83.3	92.1	90.0	101.4
			32.0	34.0	13.8	12.6	4.8	11.6	3.7	3.2
Tanzania	9.3	265.0	0.2	1.3	1.0	1.2	90.8	104.0	100.5	121.0
			18.2	46.5	29.8	35.0		30.3	12.1	13.2
Bolivia	7.6	862.3	0.2	0.7	0.6	0.7	88.8	81.3	79.4	208.6
			39.5	94.5	75.4	77.3	2.6	12.3	7.5	9.0
Ghana	6.0	302.9	0.2	0.6	0.7	0.6	102.0	82.2	99.4	105.1
			20.9	36.6	37.9	32.6	1.9	9.8	9.6	10.8
Ethiopia	6.0	89.6	0.1	1.1	0.6	1.3	85.2	100.7	83.4	119.5
			5.3	20.2	10.3	19.4		11.8	10.2	21.7
Senegal	4.9	484.9	0.2	0.6	0.5	0.4	96.5	93.5	90.2	82.3
			43.7	79.8	51.6	44.5	5.7	11.4	10.9	9.2
Mali	3.1	272.8	0.1	0.4	0.3	0.5	99.7	99.6	85.1	141.5
			18.2	45.6	31.9	41.0	7.8	15.2	13.6	15.0
Other			17.3	33.9	33.9	44.3	72.5	53.1	28.2	99.6

Notes: *OOF* Other official flows. For Israel, 1998 and 2002 are OA (official aid) flows, not ODA. Average ODA per capita is for Bangladesh (1971–73). Average ODA in per cent of GNI is for Bangladesh (1973); Bolivia (1970–73); Indonesia 1967–73); Mali (196–73); Pakistan (1967–73); Senegal (1968–73)
Source: OECD (2004)

'basic human needs' and 'redistribution with growth' were formulated. Nevertheless, the typical project aid modality remained largely unchanged.

During the 1960s and 1970s, economic progress was visible in much of the developing world. This era came to an abrupt end at the beginning of the 1980s. The international debt crisis erupted in association with macroeconomic imbalances in many countries, and it soon became evident that the downturn would be long-lasting, not temporary as in 1973. On the political scene Ronald Reagan and Margaret Thatcher came to power in the USA and UK, and at the World Bank Anne Krueger became Vice-President and Chief Economist, replacing Hollis Chenery. This change was symbolic and substantive (Kanbur 2003).

Economic circumstances in the developing countries and the relations between the North and South had changed radically. The crisis hit hard, especially in many African countries; progress over previous decades ground to a halt, inflation got out of control and the deficit in the balance payments could not be financed on a sustainable basis. Focus in development policy shifted to internal domestic failures, and achieving macroeconomic balance (externally and internally) became widely perceived as an essential prerequisite for renewed development.

'Rolling back the state' turned into a rallying call in the reform efforts, and reliance on market forces, outward orientation, and the role of the private sector, including non-governmental organizations (NGOs), was emphasized by the World Bank and others. In parallel, poverty alleviation somehow slipped out of view in mainstream agendas for economic reform, but remained at the centre of attention in more unorthodox thinking, such as the 'adjustment with a human face' approach of the UN Children's Fund (Cornia et al. 1987).

At the same time, bilateral donors and international agencies grappled with how to channel

resources to the developing world. Channelling fresh resources to developing countries in the form of discrete investment projects had become increasingly difficult. Project rates of return did not seem to justify the investments. Various kinds of quick-disbursing macroeconomic programme assistance, such as balance of payments support and sector budget support, which were not tied to investment projects and which could be justified under the headings of stabilization and adjustment, appeared to be an ideal solution to this problem. Financial programme aid and adjustment loans (and eventually debt relief) became fashionable and policy conditionality more widespread. A rationale had been found for maintaining the flow of resources, which corresponded well to the orthodox guidelines for good policy summarized by the 'Washington *consensus'* (Williamson 1997).

Meanwhile, total aid continued to grow steadily in real terms until the early 1990s, and more than tripled as a share of the growing national income of the donor community during 1970–90. After 1992, total aid flows started to decline in absolute terms (especially in the USA). Many reasons account for the fall in aggregate flows after 1992, including the decline of communism and the end of the cold war. Weakening patron–client relationships among the developing countries and the former colonial powers also played a role, and the traditional support for foreign aid by vocal interest groups in the industrial countries receded. Bilateral and multilateral aid institutions were subjected to criticism, and at times characterized as blunt instruments of commercial interests in the industrial world or as self-interested, rentseeking bureaucracies. Moreover, acute awareness in donor countries of cases of bad governance, corruption, and 'crony capitalism' led to scepticism about the credibility of governments receiving aid. Aid fatigue became widespread during the second half of the 1990s.

Aid Allocation

Foreign aid has over the years been justified in public policy pronouncements in widely differing ways, ranging from pure altruism to the shared benefits of economic development in poor countries and to the political ideology, foreign policy and commercial interests of the donor country. Few dispute that humanitarian sentiments have motivated donors. Action following severe natural calamities, which continue to be endemic in poor countries, is an example. Food and emergency relief also remains an important form of aid. In addition, the data available in Table 3 suggest that donors allocate relatively more ODA to the poorest countries. The broader validity of this casual observation is confirmed in cross-country econometric work (Alesina and Dollar 2000). While studying bilateral aid only, they conclude that most donors give more aid to poorer countries, ceteris paribus. They stress as well that there is considerable variation among donors.

Emphasis on the needs of poor countries was a prominent characteristic – and the underlying economic rationale – in much of the policy literature on foreign aid in the 1950s and 1960s. Here the focus was on estimating aid requirements in the tradition of the two-gap model (Chenery and Strout 1966). With time, development concerns have broadened. The two-gap model has become somewhat unfashionable, at least in academia, and the role of aid has changed to a much more multidimensional set of concerns (Thorbecke 2000). Nevertheless, economic development in aid-receiving countries continues as a yardstick both in its own right (at least for some donors) and as a necessary condition for the realization of other development aims.

A second observation from Table 2 is that large, populous countries, such as China and India, receive relatively small amounts of aid in per capita terms. Smaller countries such as Mali, Ghana, Bolivia and Sri Lanka are given more favourable per capita treatment. This finding is confirmed econometrically by Alesina and Dollar (2000). They stress, however, the critical and complex importance of political and strategic considerations in aid allocations.

It is not news that selfish motives are critical in donor decisions. In the past, the cold war was used as a powerful justification for providing aid to developing countries to stem the spread of communism. Similarly, aid from socialist

governments was motivated to promote socialist political and economic systems. Other strategic interests play a role as well. The USA has over the years earmarked very substantial amounts of aid to Egypt and Israel; being a former colony is an important determinant in getting access to French aid; and voting behaviour in the UN can affect aid allocation both bilaterally (Alesina and Dollar 2000) and through the multilateral system (Andersen et al. 2004).

In sum, there is often a wide gap between donor rhetoric and practice when attention is on the size and allocation of foreign aid. This gap is illustrated by the fact that the donor countries are indeed very far from contributing the 0.7 per cent of their national income as ODA, which was agreed as a UN target in 1970. As shown in Table 1, only the group of Nordic countries and the Netherlands have consistently met this target, while the USA contributed around 0.1 per cent of the US GNI in 2002. Finally, Table 3 shows that total ODA, ODA per capita, ODA as a share of GNI and ODA as a share of total flows actually vary considerably in real terms in many aid-receiving countries. Economic management in general, and management of aid inflows in particular, are not easy tasks in developing countries.

The Impact of Foreign Aid

If the economic development rationale for foreign aid is taken seriously, it is of interest to ask whether aid-receiving countries benefit from such transfers and, if so, how. What are the mechanisms through which aid works, and what are the potential negative effects associated with foreign aid? Over the past 60 years a vast amount of empirical work has (a) studied the impact of aid at micro-, meso-and macroeconomic level; (b) relied on cross-country as well as single-country data; and (c) included broad surveys of a qualitative and interdisciplinary nature as well as more strict quantitative econometric work. Many surveys are available; see for example Cassen (1987) and Tarp (2000).

An influential literature focused on cross-country econometric approaches to the analysis of aid effectiveness. This literature has gone through three generations (Hansen and Tarp 2000); and from the early 1990s macro-econometric studies came to dominate the academic and public discourse. This work was motivated in part by the availability of much better data across a range of countries and in part by insights emerging from new growth theory and the rapidly increasing number of general empirical studies of growth.

The simple Harrod–Domar model (and the two-gap Chenery–Strout extension) was used extensively in the past as the analytical framework of choice for assessing aid impact. The underlying idea was simple. Assume physical capital is the only factor of production (so investment is the key constraint on growth), and assume as well that all aid is invested. Then it is straightforward to calculate the growth impact of additional aid. If aid corresponds to six per cent of the gross national product and the capital–output ratio is estimated at 3.0, then aid adds 2.0 percentage points a year to the growth rate. The impact of aid is clearly positive, and aid works by helping to fill a savings or a foreign exchange gap.

The Achilles heel in this type of calculation is, first, that it is a tall order to expect that all aid is invested. Aid is provided for many reasons. In addition, the share of aid that ends up being invested (rather than consumed) will, in even the best of circumstances, depend on the degree of fungibility of the foreign aid transfer. Yet, even if aid adds to domestic savings and investment on less than a one-to-one basis, aid does continue to have a positive impact on growth in the traditional line of thinking – as long as total savings and investment go up.

A second line of critique of the Harrod–Domar and two-gap approach has been the argument that growth is less related to physical capital investment (including aid) than often assumed (Easterly 2001). If the key driver of the productive impact of aid is related more to incentives and relative prices and more generally to the policy environment, then it becomes important to consider potentially distortionary effects of aid on incentives and economic policies in the aid-receiving system, and vice versa. An example is 'Dutch disease', and domestic demand and resource allocation can certainly be twisted in undesirable

directions following an aid inflow. One concrete example is that aid donors often pay much higher wages to national experts and staff than equally important national institutions. Another illustration is change in the structure of domestic demand following the aid inflow.

Third, a large and growing literature on the political economy of aid (see Kanbur 2003; Gunning 2005, for references) has argued that, if aid allows a recipient government (local elites) to pursue behaviour that is in any way anti-developmental, then the potential positive impact of aid can be undermined. There are many such examples available in practice ranging from outright misuse of aid to more subtle issues such as the potential negative impact of aid on domestic taxation (Adam and O'Connell 1999).

The fear that foreign aid can generate undesirable aid dependency relationships persisted throughout the 1990s and into the 21st century, and gradually the perception that policy conditionality was failing to promote policy reform started to assert itself (Kanbur 2000; Svensson 2003). This perception prompted a keen interest in new kinds of donor–recipient relationships. One outcome was calls for increased national ownership of aid programmes. Another was that World Bank and independent academic researchers started digging into the aid–growth relationship using modern analytical techniques.

Much of the recent debate has roots in Mosley's (1987) micro–macro paradox. He suggested that, while aid seems to be effective at the microeconomic level, identifying any positive impact of aid at the macroeconomic level is harder or even impossible. Along with the implementation of adjustment programmes during the 1980s, traditional evaluation methods such as calculating the internal rate of return of projects came under severe criticism. The perception spread that aid channelled through sovereign governments is fully fungible. The internal rate of return approach also became problematic as donors started to embrace wider social goals for aid. The wave of cross-country work during the 1990s and the later, more extensive use of randomized programme evaluation (Duflo 2004) are ways of trying to come to grips with these issues.

The cross-country analysis by Boone (1996) suggested that aid does not work at all and is simply a waste of resources. This was followed up with an analysis by Burnside and Dollar (1997, 2000). They argue that some aid does work, and provided an attractive solution to the micro–macro paradox. Aid works, but only in countries with 'good policy'. They based this conclusion on an aid-policy interaction term that emerged as statistically significant in their analysis of the relationship between aid and growth.

Burnside and Dollar, and more recently Collier and Dollar (2001, 2002), have used the foregoing framework as a basis for suggesting that aid should be directed to 'good policy' countries to improve aid's impact on poverty alleviation. This recommendation is partly justified by reference to the seeming inability of aid to change policy, a finding that has emerged from other Bank-funded research (Devarajan et al. 2001). While the Bank's Monterrey document (World Bank 2002) toned down these recommendations, the basic thrust in much of the international aid debate remains that macroeconomic performance evaluation and policy criteria (established by the World Bank) should play a key role in aid allocation.

The work of Burnside, Collier, and Dollar led to discussions about what constitutes good policy. In many ways these discussions are extensions of more general debates and views about development strategy and policy, and the World Bank has gradually expanded the good policy concept to include a wider and more complex set of characteristics than originally considered. Nevertheless, if the variation in aid effectiveness across countries is not policy-induced but rather a result of poor initial conditions, a different aid allocation rule would maximize the impact of foreign aid. Moreover, the empirical finding that aid is effective, but only when accompanied by good policy, turns out to be delicate. It is robust neither to alternative specifications of the regression model (Hansen and Tarp 2001) nor to new data (Easterly et al. 2004).

Clemens et al. (2004), Dalgaard et al. (2004) and Roodman (2004) offer up-to-date accounts. It emerges that the single most common result of recent empirical studies is that aid has a positive impact on per capita growth. There is also strong

evidence to suggest that the importance of 'deep' structural characteristics is not yet fully understood. In sum, the accumulated crosscountry evidence is encouraging, and Dalgaard and Hansen (2005) estimate that the aggregate real rate of return on foreign aid financed investments is in the range of 20–25 per cent. Attention should turn to how the effectiveness of aid can and should be improved rather than concentrating on whether aid works. This implies, for example, that focus should shift from aggregate aid to different forms of aid and their application in different types of aid receiving countries – modalities matter.

Future Prospects

After many years when the project modality was the main vehicle for transferring aid, stabilization and broad structural reforms with associated programme aid were promoted vigorously in the early 1980s. A decisive shift from the state to the market as the key driver behind development was pursed. The East Asian financial crisis in 1997 signalled that the time had come for a rethink of the Washington *consensus*; and it is now widely agreed that quick-fix and single-actor approaches to development – focusing on either the state or the market – are not going to work. The state and the market have complementary roles to play in the struggle against poverty and inequality.

Aid fatigue is still evident in the international aid community, but it does seem that aid is gradually being rehabilitated from the low point of the mid-1990s. The empirical evidence that 'aid works' has been mounting steadily, and recent calls have been made for a 'big push' or a 'Marshall Plan' for Africa (World Economic Forum 2005), and foreign aid flows seem to have picked up considerably after 2002. The UN has established a target of halving world poverty by 2015 in the context of its Millennium Development Goals (MDG) (UN 2002), and the USA has embarked on a $5 billion Millennium Challenge Account (MCA) meant to stimulate aid to poor countries (Bush 2005).

All of this should not detract attention from the fact that many key challenges remain to be effectively addressed. The institutional set-up for bilateral aid delivery remains complex, uncoordinated and overburdened with many diverse tasks and aims; and calls for reform of the UN have become common. Moreover, it is far from settled where the balance between selectivity and conditionality is situated. An underlying dilemma here is that it remains disputed how the balance between real or perceived needs on the one hand and development potential and performance on the other should be struck. Various proposals and guidelines exist (including the existing IDA aid allocation formula), but much of this relies 'too heavily on a uniform model of what works in development policy' (Kanbur 2005). Past experiences provide many useful lessons about foreign aid (Robinson and Tarp 2000), but the search for more effective answers to these kinds of questions is far from complete.

Finally, aid has gradually become a much smaller player in the world economy than private capital flows. Foreign aid decision makers are well advised to try to sharpen their implementation skills and develop complementary relationships with, among others, private capital markets and NGOs (Roland-Holst and Tarp 2004). In an increasingly global world, possibilities and challenges are also opening up in the arena of international public goods. Foreign aid analysts would do well to explore these possibilities alongside more traditional investment and programme support activities, targeted on the provision of domestic public goods in poor countries.

See Also

- ▶ Development Economics
- ▶ Fiscal and Monetary Policies in Developing Countries
- ▶ International Financial Institutions (IFIs)
- ▶ Microcredit
- ▶ Third World Debt

Bibliography

Adam, C., and S. O'Connell. 1999. Aid, taxation and development in Sub-Saharan Africa. *Economics and Politics* 11: 225–254.

Alesina, A., and D. Dollar. 2000. Who gives aid to whom and why? *Journal of Economic Growth* 5: 33–63.

Andersen, T.B., T. Harr, and F. Tarp. 2004. *On US politics and IMF lending*, Discussion Paper 04–11. Copenhagen: Institute of Economics, University of Copenhagen. Online. Available at http://www.econ.ku.dk/wpa/pink/2004/0411.pdf. Accessed 22 June 2005.

Boone, P. 1996. Politics and the effectiveness of foreign aid. *European Economic Review* 40: 289–329.

Burnside, C., and D. Dollar. 1997. *Aid, policies, and growth*, Policy Research Working Paper 1777. Washington, DC: Development Research Group, World Bank.

Burnside, C., and D. Dollar. 2000. Aid, policies, and growth. *American Economic Review* 90: 847–868.

Bush, G. 2005. The millennium challenge account. Online. Available at http://www.whitehouse.gov/infocus/developingnations/millennium.html. Accessed 22 June 2005.

Cassen, R., et al. 1987. *Does aid work?* Oxford: Clarendon Press.

Chenery, H., and A.M. Strout. 1966. Foreign assistance and economic development. *American Economic Review* 56: 679–733.

Clemens, M., S. Radelet, and R. Bhavnani. 2004. *Counting chickens when they hatch: The short-term effect of aid on growth*, Working Paper 44. Washington, DC: Center for Global Development. Online. Available at http://www.cgdev.org/ Publications/index.cfm?PubID=130. Accessed 22 June 2005.

Collier, P., and D. Dollar. 2001. Can the world cut poverty in half? How policy reform and effective aid can meet the international development goals. *World Development* 29: 1787–1802.

Collier, P., and D. Dollar. 2002. Aid allocation and poverty reduction. *European Economic Review* 46: 1475–1500.

Cornia, G., R. Jolly, and F. Stewart. 1987. *Adjustment with a human face: Protecting the vulnerable and promoting growth*. Oxford: Clarendon Press.

Dalgaard, C.-J., and H. Hansen. 2005. *The return to foreign aid*, Discussion Paper 05–04. Copenhagen: Institute of Economics, University of Copenhagen. Online. Available at http://www.econ.ku.dk/wpa/pink/2005/0504.pdf. Accessed 22 June 2005.

Dalgaard, C.-J., H. Hansen, and F. Tarp. 2004. On the empirics of foreign aid and growth. *Economic Journal* 114: F191–F216.

Devarajan, S., D. Dollar, and T. Holmgren, eds. 2001. *Aid and reform in Africa*. Oxford: Oxford University Press for the World Bank.

Duflo, E. 2004. Evaluating the impact of development aid programs: The role of randomized evaluation. Paper presented at the 2nd AFD-EUDN Conference, Paris, 25 November. Online. Available at http://www.eudnet.net/Download/AfD-EUDN04_Duflo.pdf. Accessed 22 June 2005.

Easterly, W. 2001. *The elusive quest for growth*. Cambridge, MA: MIT Press.

Easterly, W., R. Levine, and D. Roodman. 2004. Aid, policies, and growth: Comment. *American Economic Review* 94: 774–780.

Gunning, J. 2005. Why give aid? Original paper presented at the 2nd AFD-EUDN Conference, Paris, 25 November. Revised paper online. Available at http://www.eudnet.net/Download/AfD-EUDN04_Gunning-revised.pdf. Accessed 22 June 2005.

Hansen, H., and F. Tarp. 2000. Aid effectiveness disputed. *Journal of International Development* 12: 375–398.

Hansen, H., and F. Tarp. 2001. Aid and growth regressions. *Journal of Development Economics* 64: 547–570.

Hjertholm, P., and H. White. 2000. Foreign aid in historical perspective: Background and trends. In *Foreign aid and development: Lessons learnt and directions for the future*, ed. F. Tarp. London: Routledge.

Kanbur, R. 2000. Aid, conditionality and debt in Africa. In *Foreign aid and development: Lessons learnt and directions for the future*, ed. F. Tarp. London: Routledge.

Kanbur, R. 2003. *The economics of international aid*, Working Paper 39. New York: Department of Applied Economics and Management, Cornell University. Online. Available at http://www.arts.cornell.edu/poverty/kanbur/HandbookAid.pdf. Accessed 22 June 2005.

Kanbur, R. 2005. Reforming the formula: A modest proposal for introducing development outcomes in IDA allocation procedures. Original paper presented at the 2nd AFD-EUDN Conference, Paris, 25 November. Revised paper online. Available at http://www.arts.cornell.edu/poverty/kanbur/IDAForm.pdf. Accessed 22 June 2005.

Mosley, P. 1987. *Overseas aid: Its defense and Reform*. Brighton: Wheatsheaf Books.

OECD (Organization for Economic Development and Cooperation). 2004. Development Assistance Committee: Disbursements and commitments of official and private flows (Table 1) and Development Assistance Committee: destination of official development assistance and official aid – disbursements (Table 2a). Online. Available at http://www.oecd.org/dac/stats/idsonline. Accessed 16 Feb 2005.

Robinson, S., and F. Tarp. 2000. Foreign aid and development: summary and synthesis. In *Foreign aid and development: Lessons learnt and directions for the future*, ed. F. Tarp. London: Routledge.

Roland-Holst, D., and F. Tarp. 2004. New perspectives on aid effectiveness. In *Toward pro-poor policies – aid, institutions and globalization*, ed. B. Tungodden, N. Stern, and I. Kolstad. Washington, DC: World Bank and Oxford University Press.

Roodman, D. 2004. *The anarchy of numbers: Aid, development, and cross-country empirics*, Working paper 32. Washington, DC: Center for Global Development. Online. Available at http://www.cgdev.org/Publications/?PubID=36. Accessed 22 June 2005.

Rosenstein-Rodan, P. 1943. Problems of industrialization of Eastern and South-Eastern Europe. *Economic Journal* 53: 202–211.

Svensson, J. 2003. Why conditional aid doesn't work and what can be done about it? *Journal of Development Economics* 70: 381–402.
Tarp, F., ed. 2000. *Foreign aid and development: Lessons learnt and directions for the future*. London: Routledge.
Thorbecke, E. 2000. The evolution of the development doctrine and the role of foreign aid, 1950–2000. In *Foreign aid and development: Lessons learnt and directions for the future*, ed. F. Tarp. London: Routledge.
UN (United Nations). 2002. The millennium development goals and the United Nations role. Fact sheet. Online. Available at http://www.un.org/millenniumgoals/MDGs-FACTSHEET1.pdf. Accessed 16 Feb 2005.
Williamson, J. 1997. The Washington Consensus revisited. In *Economic and social development into the XXI Century*, ed. L. Emmerij. Washington, DC: Inter-American Development Bank.
World Bank. 2002. *A case for aid: Building a consensus for development assistance*. Washington, DC: World Bank.
World Economic Forum. 2005. Global leaders call for big push on aid to Africa at World Economic Forum meeting in Davos. Press Release. Online. Available at http://www.csrwire.com/article.cgi/3480.html. Accessed 22 June 2005.

Foreign Direct Investment

Bruce A. Blonigen

Abstract

Foreign direct investment (FDI) occurs when an individual or firm acquires controlling interest in productive assets of another country. We review the literature on FDI, which can be divided into two broad categories. The first is the inquiry into why multinational production occurs and the factors that determine the patterns of worldwide FDI. The second is the impact that FDI and multinational enterprises (MNEs) have on the parent and host countries, including economic growth, returns to factors of production, and externalities.

Keywords

Agglomeration externalities; Double taxation issue; Efficiency wages; Exchange rate volatility; Factor endowments; Firm, theory of; Foreign direct investment; General equilibrium model; Greenfield foreign direct investment; Horizontal foreign direct investment; Information externalities; International capital flows; Intra-firm trade; Knowledge-capital model of multinational enterprises; Multinational enterprises (MNEs); Multinational firms; Ownership-location-internalization' (OLI) theory of multinational enterprises; Partial equilibrium model of firm behaviour; Portfolio investment; Productivity spillovers; Size of nations; Tax competition; Tax treaties; Taxation of corporate profits; Trade protection; Transactions costs; Vertical foreign direct investment; Wage heterogeneity, sources of; Wage spillovers

JEL Classifications

F21; F23

Foreign direct investment (FDI) occurs when an individual or firm acquires a controlling interest (typically defined as at least ten per cent ownership) in productive assets in another country. This contrasts with portfolio investment, which includes purchases of foreign bonds, currencies, and stocks in amounts that do not provide control. The most common method of FDI is through the acquisition of a firm. Construction of a new plant is also common and typically referred to as 'greenfield' FDI. Other forms of FDI include partnerships in a foreign joint venture and earnings reinvested in an existing foreign affiliate. Firms with affiliates in more than one country are termed 'multinational enterprises' (MNEs).

While real world GDP grew at a 2.5 per cent annual rate and real world exports grew by 5.6 per cent annually from 1986 through 1999, real world FDI inflows grew by 17.7 per cent over this same period (Giorgio and Venables 2004). Additionally, Bernard et al. (2005) find that 90 per cent of US exports and imports flow through MNEs, with roughly 50 per cent of US trade flows occurring between affiliates of the same MNE, or what is termed 'intra-firm trade'. While the majority of FDI flows are between developed countries, FDI accounted for the majority of capital flows to

less-developed countries from 1990 to 2003 (UNCTAD 2004).

The study of FDI can be divided into two broad categories. The first is the inquiry into why multinational production occurs and the factors that determine the patterns of worldwide FDI. The second is the impact that FDI and MNEs have on the parent and host countries, including economic growth, returns to factors of production, and externalities for innovative activity.

Understanding What Motivates FDI by MNEs

Theory

Theoretical treatment of FDI and MNEs in the economics profession can be traced back to the 1970s, when researchers began to consider why some firms choose to locate production abroad rather than serve such markets through exports or licensing. A key insight is that MNEs may be distinguished by their ownership of firm-specific assets for which market failures can make exporting or licensing arrangements less attractive to the firm than FDI. For example, a foreign licensee may not offer full value in negotiations over a contract if the firm-specific asset is intangible and not fully revealed (for example, a unique production process), but the licensor firm will not want to reveal the asset fully until a contract is finalized. The costs associated with this inherent hold-up problem may then lead the firm to set up its own affiliate in the foreign market. This is termed 'internalization' in the literature, and forms the key element in the 'ownership-location-internalization' (OLI) theory of MNEs that developed out of this era and has been surveyed recently by Dunning (2001).

The OLI theory is an international business concept that was never formally represented in a mathematical model. As such, the international economics literature continued to treat FDI as simply another capital flow until the mid-1980s, even though its features and patterns differed from those of other capital flows. This changed with papers by Markusen (1984) and Helpman (1984) that developed general equilibrium models of MNEs. Both papers focused on another feature of firm-specific assets, namely, the public-goods aspect of many firm-specific assets that can be applied simultaneously in production across all plants owned by the firm. This feature of firm-specific assets makes it more attractive for a firm to build multiple plants, though something else must be added to a model to explain locating plants into foreign countries. In Helpman (1984) this is accomplished by assuming that MNEs can be separated into two types of activities: a skill-intensive headquarters that generates the firm-specific assets, and a low-skill-intensive production process. If endowment differences are sufficient across countries, MNEs will vertically separate the firm between headquarter services in the skill-abundant parent country and production in the low-skill host country. This type of model is called a 'vertical FDI' model. In contrast, Markusen's (1984) model generates multiplant MNEs through the introduction of trade costs (that is, transportation costs, trade barriers, and so on) that are large enough that an MNE chooses to replicate itself in the foreign country to serve the market there. This type of model is termed 'horizontal FDI'.

These models have become the main theoretical MNE frameworks for trade economists, as recent literature has extended these models. Brainard (1997) develops and tests hypotheses from a simplified horizontal MNE model assuming monopolistic competition. Markusen et al. (1996) develop an MNE model that blends both the horizontal and vertical models into what is termed the 'knowledge-capital' model. More recently, Helpman et al. (2004) have developed a model that can explain the coexistence of both exporting and MNEs in the same industry by allowing for heterogeneity across firms; other papers have developed models that formalize the role of transactions costs and theory of the firm (for example, Antras and Helpman 2004; Feenstra and Hanson 2005).

Empirics

Empirical work on the factors that determine FDI patterns has focused primarily on the effect of government policies and macroeconomic

phenomena such as exchange rates and taxes. Most of these studies motivate their analyses with a partial equilibrium model of firm behaviour responding to these various factors. Only recently have empirical studies examined the more fundamental long-run drivers of total FDI activity, such as country size and factor endowments, as predicted by the general equilibrium modelling discussed above. Availability of micro-level data has been an issue for the literature as well. Testing theories of firm-level models with industry-or country-level data requires strong assumptions about firm characteristics. While firm-level data is being employed more often in recent work, much of the literature has examined more aggregate data.

Exchange Rates

The effects of exchange rate movements on FDI are not immediately obvious. If a host country's currency depreciates relative to the parent country's currency, this lowers the price of host-country assets. However, if the asset generates returns in the host country's currency, these returns have likewise depreciated in the parent-country currency. Froot and Stein (1991) and Blonigen (1997), however, provide theoretical links that predict that host-country depreciations increase inbound FDI; and empirical evidence generally supports this. A related literature has examined how exchange rate expectations may affect FDI decisions. Campa (1993) provides theory and evidence that exchange rate uncertainty will decrease FDI, while Cushman (1985) and Goldberg and Kolstad (1995) conclude that quite opposite results can be expected and found depending on the firm's trade linkages across markets. On a final note, there has been recent work on the impact of exchange rate crises on FDI. Surprisingly, FDI is relatively stable through currency crises in host countries and, in fact, Aguiar and Gopinath (2005) show that MNEs opportunistically increase their investments in these host countries.

Taxes

Like exchange rate movements, the effect of taxes on FDI has not proven to be straightforward either. While there is an array of taxes that may affect FDI, the primary focus has been on corporate income tax rates in host countries. The natural hypothesis is that higher host-country tax rates discourage FDI, and a survey by de Mooij and Ederveen (2003) finds a median elasticity of tax rates on FDI of minus 3.3 across 25 different empirical studies. However, the literature has also shown that the effects of taxes on FDI can vary substantially depending on the type of taxes, the form of FDI (see, for example, Hartman 1985), and the influence of government policy.

Perhaps the most explored issue in this literature has been the issue of how parent countries deal with the 'double taxation' issue – taxation in both host and parent countries. The common distinction is between territorial countries that do not tax any income outside of the parent country, exempting foreign-earned income from tax liability, and a worldwide tax method which considers all earned income by its parent firms potentially taxable, but may treat foreign income in a number of ways to avoid double taxation of the MNE. The standard treatment to deal with this double taxation issue is for the home country to offer a credit or a deduction of foreign tax payment made by the MNE. A number of studies of the US 1986 tax reform find mixed evidence for differences in FDI behaviour under different parent-country tax regimes (for example, Scholes and Wolfson 1990; Swenson 1994). Much stronger results come from work by Hines (1996) which finds that US taxation decreases FDI more for non-credit-system foreign investors than for credit-system foreign investors.

A final significant literature in this area is tax competition between countries competing for FDI (for example, Janeba 1995) and the impact of bilateral tax treaties between countries (for example, Chisik and Davies 2004). Hines (1999) and Gresik (2001) have excellent surveys of the FDI and taxation literature.

Other Factors

A variety of other smaller literatures have investigated the effect of other factors on FDI. These include the effects of host-country institutions (Wei 2000), trade protection policies, and

agglomeration and information externalities (Head et al. 1995; Blonigen et al. 2005).

Examination of General-Equilibrium Model Predictions

More recently, empirical efforts have been made to more closely match empirical specifications of country-level FDI activity with general-equilibrium models of MNEs. Most previous empirical work uses gravity-based variations to model country-level FDI patterns where size of countries and distance between them are key regressors. Carr et al. (2001) instead lay out an empirical specification based on the knowledge-capital model of MNE activity which suggests that factor endowment differences are an important control not found in gravity-based specifications. These endowment differences are important as they proxy for vertical MNE motivations. While Carr et al. (2001) find that the data fit the knowledge-capital model, follow-up work has found specification issues that call into question evidence of vertical motivations for FDI (see Blonigen et al. 2003; Braconier et al. 2005). Alternative approaches by Yeaple (2003b) and Hanson et al. (2005), however, have confirmed vertical motivations in the data, at least for certain sectors such as electronics and transportation equipment. Another concern pointed out by Yeaple (2003a) is that third country interactions may matter for FDI patterns. Recent empirical work by Baltagi et al. (2007) suggests that such effects are important empirically.

The Economic Impact of FDI and MNE Activity

A second significant part of the FDI literature is the examination of FDI impacts on parent and, particularly, host countries. The primary areas of study have been on the effect of FDI on host country wages, technology spillovers, and economic growth.

Studies of FDI effects on host-country wages typically begin with the hypothesis that MNEs raise wages in the host country. Part of this is ascribed to the fact that the value of marginal product will be higher with MNEs due to productivity advantages and, thus, MNEs pay higher wages. However, an argument can also be made that MNEs need to pay higher efficiency wages than local firms to attract quality workers in an environment which they are relatively uninformed. Regardless of the explanation, the empirical evidence clearly shows that MNEs pay higher wages in both developed countries (for example, Globerman et al. 1994) and less-developed ones (for example, Aitken et al. 1996).

The more intriguing question is whether there are wage spillovers, in the sense that MNEs raise the wages paid by local firms as well. Spillovers are inherently difficult to identify in the data. Virtually all of the studies rely on interpreting a positive correlation between the presence of foreign firms in a local industry and the wages of local firms as evidence of spillovers. Not surprisingly, the evidence is decidedly mixed across numerous studies, as discussed by Lipsey and Sjöholm (2005). The theoretical development behind this issue is also relatively undeveloped in the literature as to when and where we should expect such wage spillovers.

A related issue is the effects of FDI on wage inequality. If MNEs have different technologies that demand different types of labour from local firms, increased FDI can lessen or exacerbate existing wage inequality. There are a number of cross-country studies that find a variety of FDI effects on wage inequality for the host country. Results for the United States using more detailed industry-level data likewise indicate little to no impact of outbound or inbound FDI on US wage inequality (Slaughter 2000; Blonigen and Slaughter 2001). Feenstra and Hanson (1997) provides a model to show how FDI can increase the difference between skilled and unskilled workers' wages in both host and parent countries with empirical work that finds strong impacts of US FDI activity on Mexican wage inequality.

The literature on productivity spillovers from FDI is vast compared with the one on wage spillovers, yet the evidence is decidedly mixed as well (see Görg and Strobl 2001, for a survey). This is not surprising in many ways. First, theory is ambiguous on this issue. Foreign firms are presumably more

efficient than the average local firm. Thus, FDI lowers market shares for local firms, which can lead to productivity losses for these firms, particularly if economies of scale are important. However, better technologies of foreign firms may ultimately leak to local firms through, for example, former employees or common suppliers. The second likely reason for mixed evidence is again the difficulty of identifying spillovers in the data (see Aitken and Harrison 1999, for a discussion).

There is also a significant literature that attempts to gauge the overall impact of FDI on a host economy's economic growth. Like the trade and growth literature, this is difficult because of the obvious endogeneity issue, which is difficult to overcome. Such a question also relies on aggregate cross-country data, which is often quite poor. Most papers in the literature do not adequately control for these issues, and Carkovic and Levine (2005) points out the statistical sensitivity of these studies' results.

There are much smaller literatures on a variety of other host- and parent-country effects of FDI. This includes the impact of FDI on parent-country investment and employment (Blomström et al. 1997), the effects of FDI on host-country trade policies (Blonigen and Figlio 1998), and differences in how MNEs adjust to local factor prices (Giorgio et al. 2003).

See Also

- ▶ International Capital Flows
- ▶ Location Theory

Bibliography

Aguiar, M., and G. Gopinath. 2005. Fire-sale foreign direct investment and liquidity crises. *Review of Economics and Statistics* 87: 439–452.

Aitken, B., and A. Harrison. 1999. Do domestic firms benefit from direct foreign investment? Evidence from Venezuela. *American Economic Review* 89: 605–618.

Aitken, B., A. Harrison, and R. Lipsey. 1996. Wages and foreign ownership: A comparative study of Mexico, Venezuela, and the United States. *Journal of International Economics* 40: 345–371.

Antras, P., and E. Helpman. 2004. Global sourcing. *Journal of Political Economy* 112: 552–580.

Baltagi, B., P. Egger, and M. Pfaffermayr. 2007. Estimating models of complex FDI: Are there third-country effects? *Journal of Econometrics* 140(1): 5–51.

Bernard, A., J. Jensen, and P. Schott. 2005. Importers, exporters and multinationals: A portrait of the firms in the U.S. that trade goods. Working paper no. 11404. Cambridge, MA: NBER.

Blomström, M., G. Fors, and R. Lipsey. 1997. Foreign direct investment and employment: Home country experience in the United States and Sweden. *Economic Journal* 107: 1787–1797.

Blonigen, B. 1997. Firm-specific assets and the link between exchange rates and foreign direct investment. *American Economic Review* 87: 447–465.

Blonigen, B., and D. Figlio. 1998. Voting for protection: Does direct foreign investment influence legislator behavior? *American Economic Review* 88: 1002–1014.

Blonigen, B., and M. Slaughter. 2001. Foreign-affiliate activity and U.S. skill upgrading. *Review of Economics and Statistics* 83: 362–376.

Blonigen, B., R. Davies, and K. Head. 2003. Estimating the knowledge-capital model of the multinational enterprise: Comment. *American Economic Review* 93: 980–994.

Blonigen, B., C. Ellis, and D. Fausten. 2005. Industrial groupings and foreign direct investment. *Journal of International Economics* 65: 75–91.

Braconier, H., P.-J. Norback, and D. Urban. 2005. Reconciling the evidence on the knowledge-capital model. *Review of International Economics* 13: 770–786.

Brainard, S. 1997. An empirical assessment of the proximity-concentration trade-off between multinational sales and trade. *American Economic Review* 87: 520–544.

Campa, J. 1993. Entry by foreign firms in the U.S. under exchange rate uncertainty. *Review of Economics and Statistics* 75: 614–622.

Carkovic, M., and R. Levine. 2005. Does foreign direct investment accelerate economic growth? In *Does foreign direct investment promote development?* ed. T. Moran, E. Graham, and M. Blomström. Washington, DC: Institute for International Economics.

Carr, D., J. Markusen, and K. Maskus. 2001. Estimating the knowledge-capital model of the multinational enterprise. *American Economic Review* 91: 693–708.

Chisik, R., and R. Davies. 2004. Asymmetric FDI and tax-treaty bargaining: Theory and evidence. *Journal of Public Economics* 88: 1119–1148.

Cushman, D. 1985. Real exchange rate risk, expectations, and the level of direct investment. *Review of Economics and Statistics* 67: 297–308.

de Mooij, R., and S. Ederveen. 2003. Taxation and foreign direct investment: A synthesis of empirical research. *International Tax and Public Finance* 10: 673–693.

Dunning, J. 2001. The eclectic (OLI) paradigm of international production: Past, present and future. *International Journal of Economics and Business* 8: 173–190.

Feenstra, R., and G. Hanson. 1997. Foreign direct investment and relative wages: Evidence from Mexico's maquiladoras. *Journal of International Economics* 42: 371–393.

Feenstra, R., and G. Hanson. 2005. Ownership and control in outsourcing to China: Estimating the property-rights theory of the firm. *Quarterly Journal of Economics* 120: 729–761.

Froot, K., and J. Stein. 1991. Exchange rates and foreign direct investment: An imperfect capital markets approach. *Quarterly Journal of Economics* 106: 1191–1217.

Giorgio, B., and A. Venables. 2004. *Multinational firms in the world economy*. Princeton/Oxford: Princeton University Press.

Giorgio, B., D. Checci, and A. Turrini. 2003. Adjusting labor demand: Multinational versus national firms: A cross-European analysis. *Journal of the European Economic Association* 1: 708–719.

Globerman, S., J. Ries, and I. Vertinsky. 1994. The economic performance of foreign affiliates in Canada. *Canadian Journal of Economics* 27: 143–156.

Goldberg, L., and C. Kolstad. 1995. Foreign direct investment, exchange rate variability and demand uncertainty. *International Economic Review* 36: 855–873.

Görg, H., and E. Strobl. 2001. Multinational companies and productivity spillovers: A meta-analysis. *Economic Journal* 111: 723–739.

Gresik, T. 2001. The taxing task of taxing transnationals. *Journal of Economic Literature* 39: 800–838.

Hanson, G., R. Mataloni, and M. Slaughter. 2005. Vertical production networks in multinational firms. *Review of Economics and Statistics* 87: 664–678.

Hartman, D. 1985. Tax policy and foreign direct investment. *Journal of Public Economics* 26: 107–121.

Head, K., J. Ries, and D. Swenson. 1995. Agglomeration benefits and location choice: Evidence from Japanese manufacturing investments in the United States. *Journal of International Economics* 38: 223–247.

Helpman, E. 1984. A simple theory of international trade with multinational corporations. *Journal of Political Economy* 92: 451–471.

Helpman, E., M. Melitz, and S. Yeaple. 2004. Export versus FDI with heterogeneous firms. *American Economic Review* 94: 300–316.

Hines Jr., J. 1996. Altered states: Taxes and the location of foreign direct investment in America. *American Economic Review* 86: 1076–1094.

Hines Jr., J. 1999. Lessons from behavioral responses to international taxation. *National Tax Journal* 52: 305–322.

Janeba, E. 1995. Corporate income tax competition, double taxation treaties, and foreign direct investment. *Journal of Public Economics* 56: 311–325.

Lipsey, R., and F. Sjöholm. 2005. The impact of inward FDI on host countries: Why such different answers? In *Does foreign direct investment promote development?* ed. T. Moran, E. Graham, and M. Blomström. Washington, DC: Institute for International Economics.

Markusen, J. 1984. Multinationals, multi-plant economies, and the gains from trade. *Journal of International Economics* 16: 205–226.

Markusen, J., A. Venables, D. Eby-Konan, and K. Zhang. 1996. A unified treatment of horizontal direct investment, vertical direct investment and the pattern of trade in goods and services. Working paper no. 5696. Cambridge, MA: NBER.

Scholes, M., and M. Wolfson. 1990. The effects of changes in tax law on corporate reorganization activity. *Journal of Business* 63: S141–S164.

Slaughter, M. 2000. Production transfer within multinational enterprises and American wages. *Journal of International Economics* 50: 449–472.

Swenson, D. 1994. The impact of U.S. tax reform on foreign direct investment in the United States. *Journal of Public Economics* 54: 243–266.

UNCTAD (United Nations Conference on Trade and Development). 2004. *World investment report 2004: The shift to services*. New York/Geneva: United Nations.

Wei, S.-J. 2000. How taxing is corruption on international investors? *Review of Economics and Statistics* 82: 1–11.

Yeaple, S. 2003a. The complex integration strategies of multinationals and cross country dependencies in the structure of foreign direct investment. *Journal of International Economics* 60: 293–314.

Yeaple, S. 2003b. The role of skill endowments in the structure of U.S. outward foreign direct investment. *Review of Economics and Statistics* 85: 726–734.

Foreign Exchange Market Microstructure

Martin D. D. Evans

Abstract

Research on foreign exchange market microstructure focuses on the idea that trading is an integral part of the process whereby information relevant to the pricing of foreign currency becomes embedded in spot rates. Micro-based models of this process produce empirical predictions that find strong support in the data. Micro-based models can account for a large proportion of the daily variation in spot rates. They also supply a rationale for the apparent disconnect between spot rates and fundamentals. Micro-based models provide

out-of-sample forecasting power for spot rates that is an order of magnitude above that usually found in exchange-rate models.

Keywords

Arbitrage; Common knowledge news; Depreciation rates; Exchange rate dynamics; Exchange rate puzzles; Financial market contagion; Foreign exchange market microstructure; Foreign exchange risk premium; Information aggregation; Order flows; Spot exchange rates; Stop-loss orders

JEL Classifications

F3; F33

Models of foreign exchange (FX) market microstructure examine the determination and behaviour of spot exchange rates in an environment that replicates the key features of trading in the FX market. Traditional macro exchange-rate models pay little attention to how trading in the FX market actually takes place. The implicit assumption is that the details of trading (that is, who quotes currency prices and how trade takes place) are unimportant for the behaviour of exchange rates over months, quarters or longer. Micro-based models, by contrast, examine how information relevant to the pricing of foreign currency becomes reflected in the spot exchange rate via the trading process. According to this view, trading is not an ancillary market activity that can be ignored when one considers exchange rate behaviour. Rather, trading is an integral part of the process through which spot rates are determined and evolve. Recent micro-based FX models also differ from other areas of microstructure research in their focus on the links between trading, asset price dynamics and the macroeconomy.

Recent research on exchange rates stresses the role of heterogeneity (for example, Bacchetta and van Wincoop 2006; Hau and Rey 2006). Micro-based exchange-rate models start from the premise that much of the information about the current and future state of the economy is dispersed across agents (that is, individuals, firms and financial institutions). Agents use this information in making their everyday decisions, including decisions to trade in the FX market at the prices quoted by dealers. Dealers quote prices (for example, dollars per unit of foreign currency) at which they stand ready to buy or sell foreign currency; they will purchase foreign currency at their bid quote, and sell foreign currency at their ask quote. Agents that choose to trade with an individual dealer are termed the 'dealer's customers'. The difference between the value of purchase and sale orders *initiated* by customers during any trading period is termed 'customer order flow'. Importantly, order flow is different from trading volume because it conveys information. Positive (negative) order flow indicates to dealers that, on balance, their customers value foreign currency more (less) than their asking (bid) price. By tracking who initiates each trade, order flow provides a measure of the information exchanged between counterparties in a series of financial transactions.

Trading in the FX market also takes place between dealers. In direct inter-dealer trading, one dealer asks another for a bid and ask quote, and then decides whether he wishes to trade. When the dealer initiating the trade purchases (sells) foreign currency, the trade generates a positive (negative) inter-dealer order flow equal to the value of the purchase (sale). Inter-dealer trading can also take place indirectly via brokerages that act as intermediaries between two or more dealers. In recent years electronic brokerages have come to dominate inter-dealer trading, but the inter-dealer order flow generated by brokered trades plays the same informational role as the order flow associated with direct inter-dealer trading.

Micro-Based Exchange Rate Determination

At first sight, the pattern of FX trading activity seems far too complex to provide any useful insight into the behaviour of exchange rates. However, on closer examination two key features emerge. First, the equilibrium spot exchange rate does not come out of a 'black box'. Instead, it is solely a function of the foreign currency prices quoted by dealers at a point in time. This is a

distinguishing feature of micro-based exchange rate models and has far-reaching implications. Second, information about the current and future state of the economy will impact on exchange rates only when, and if, it affects dealer quotes. Dealers may revise their quotes in response to new public information that arrives via macroeconomic announcements. They may also revise their quotes based on orders they receive from customers and other dealers. This order flow channel is the means though which dispersed information concerning the economy affects dealer quotes and hence the spot exchange rate. The role played by order flow in transmitting information to dealers, and hence to their quotes, is another distinguishing feature of micro-based exchange rate models.

Micro-based models incorporate these two features of FX trading into a simplified setting. Canonical multi-dealer models, such as Lyons (1997) and Evans and Lyons (2002a), posit a simple sequence of quoting and trading. At the start of each period, dealers quote FX prices to customers. These prices are assumed to be good for any amount and are publicly observed. Each dealer then receives orders from a subset of agents, his customers. Dealers next quote prices in the inter-dealer market. These prices, too, are good for any quantity and are publicly observed. Dealers then have the opportunity to trade among themselves. Inter-dealer trading is simultaneous and trading with multiple partners is feasible.

In this trading environment, optimal quote decisions take a simple form; all dealers quote the same FX price to both customers and other dealers. We can represent the period-t quote as

$$s_t = (1-b)\sum_{i=0}^{\infty} b^i E\left[f_{t+i} \middle| \Omega_j^D\right], \qquad (1)$$

where $0 < b < 1$. s_t is the log price of foreign currency quoted by all dealers, and f_t denotes exchange rate fundamentals. The form for fundamentals differs according to the macroeconomic structure of the model. For example, in Evans and Lyons (2004b) f_t includes home and foreign money supplies and household consumption. In models where central banks conduct monetary policy via the control of short-term interest rates (that is, follow Taylor rules), f_t will include variables used to set policy. More generally, f_t will include a term that identifies the foreign exchange risk premium.

While Eq. (1) takes the present value form familiar from standard international macro models, here it represents how dealers quote the price for foreign currency in equilibrium. All dealers choose to quote the same price in this trading environment because doing otherwise opens them up to arbitrage, a costly proposition. (Recall that quotes are publicly observed and good for any amount, so any discrepancy between quotes would represent an opportunity for a riskless trading profit.) Consequently, the month-t quote must be a function of information known to all dealers. Equation (1) incorporates this requirement with the use of the expectations operator, $E\left[\cdot \middle| \Omega_t^D\right]$, that denotes expectations conditioned on information common to all dealers at the start of month t, Ω_t^D. This is not to say that all dealers have the same information. On the contrary, the customer order flows received by individual dealers represent an important source of private information, so there may be a good deal of information heterogeneity across dealers at any one time. The important point to note from Eq. (1) is that, due to the 'fear of arbitrage', individual dealers choose not to quote prices based on their own private information. In this trading environment dealers use their private information in initiating trade with other dealers, and, in so doing, contribute to the process through which all dealers acquire information.

The implications of micro-based models for the dynamics of spot rates are most easily seen by rewriting (1) as

$$\Delta s_{t+1} = \frac{1-b}{b}\left(s_t - E\left[f_t \middle| \Omega_j^D\right]\right) + \varepsilon_{t+1}, \qquad (2)$$

where $\Delta s_{t+1} = s_{t+1} - s_t$, and

$$\varepsilon_{t+1} = \frac{1-b}{b}\sum_{i=1}^{\infty} b^i \left(E\left[f_{t+i} \middle| \Omega_{t+1}^D\right] - E\left[f_{t+i} \middle| \Omega_t^D\right]\right), \qquad (3)$$

Equation (2) decomposes the change in the log spot rate (that is, the depreciation rate for the home currency) into two components: the expected change $E\left[\Delta s_{t+1} | \Omega_t^D\right]$ identified by the first term, and the unexpected change, $\varepsilon_{t+1} = s_{t+1} - E\left[s_{t+1} | \Omega_t^D\right]$, shown in Eq. (3). Both terms contribute to exchange rate dynamics in micro-based models. In equilibrium, dealers' period-t quote must be based on expectations, $E\left[\Delta s_{t+1} | \Omega_t^D\right]$, that match the risk-adjusted returns on different assets. This means that variations in the interest differential between home and foreign bonds can contribute to the volatility of the depreciation rate via the first term in (2). The second term, ε_{t+1}, identifies the impact of new information received by all dealers between the start of periods t and $t + 1$. Equation (3) shows that new information impacts on the FX price quoted in period $t + 1$ to the extent that it revises forecasts of the present value of fundamentals based on dealers' common information.

As an empirical matter, depreciation rates are very hard to forecast, so the dynamics of spot rates are largely attributable to the effects of news. Here micro-based models have a big advantage over their traditional counterparts because their trade-based foundations provide detail on how news affects spot rates. In particular, as Eq. (3) indicates, micro-based models focus on how new information about the fundamentals reaches dealers and induces them to revise their FX quotes.

News concerning fundamentals can reach dealers either directly or indirectly. Common knowledge (CK) news operates via the direct channel. CK news contains unambiguous information about current and/or future fundamentals that is simultaneously observed by all dealers and immediately incorporated into the FX price they quote. In principle, macroeconomic announcements (for example, on GDP, industrial production or unemployment) could be a source for CK news, but in practice they rarely contain much unambiguous new information. In fact, CK news events appear rather rare. The indirect channel operates via order flow and conveys dispersed information about fundamentals to dealers. Dispersed information comprises micro-level information on economic activity that is correlated with fundamentals. Examples include the sales and orders for the products of individual firms, market research on consumer spending, and private research on the economy conducted by financial institutions. Dispersed information first reaches the FX market via the customer order flows received by individual dealers. These order flows have no immediate impact on dealer quotes because they represent private information to the recipient dealer. The information in each customer flow will impact on quotes only when it is known to all dealers. Inter-dealer order flow is central to this process. Individual dealers use their private information to trade in the inter-dealer market. In so doing, information on their customer orders is aggregated and spread across the market. This process is known as 'information aggregation'. Dispersed information is incorporated into dealer quotes once this process is complete.

Empirical Evidence

The appeal of micro-based models is not solely based on their theoretical foundations. In marked contrast with traditional exchange-rate models, micro-based models have enjoyed a good deal of empirical success. Evans and Lyons (2002a) first demonstrated their empirical power when studying the relation between depreciation rates and inter-dealer order flow at the daily frequency. In particular, they show that aggregate inter-dealer order flow from trading in the spot dollar/deutschmark market on day d accounts for 64 per cent of the variation in the depreciation rate, Δs_{d+1}, between the start of days d and $d + 1$. This is a striking result because macro models can account for less than one per cent of daily depreciation rates. It is also readily explained in terms of Eqs. (2) and (3). Aggregate inter-dealer order flow during day d trading provides a measure of the market-wide information flow that dealers use to revise their quotes between the start of days d and $d + 1$. This contemporaneous relationship between depreciation rates and inter-dealer order flows appears robust. It holds for many different currencies and for different currency-order flow combinations

(for example, Evans and Lyons 2002b; Payne 2003; Froot and Ramadorai 2005). It is also worth emphasizing that order flow's impact on spot rates is very persistent. There is very little serial correlation in the daily depreciation rates for major currencies, so the order flow impact on current FX quotes persists far into the future.

While consistent with the idea that dispersed information is impounded into spot exchange rates via inter-dealer order flow, these results do not provide direct evidence on the ultimate source of exchange rate dynamics. According to micro-based models, the analysis of customer order flows should provide the evidence. In particular, if inter-dealer order flows measure the market-wide flow of information concerning fundamentals originally motivating customer orders, customer orders should also have explanatory power for depreciation rates. This is indeed the case. Evans and Lyons (2004b) show that a significant contemporaneous relationship exists between depreciation rates and the customer order flows of a single large bank. Moreover, the strength of this relationship increases as we move from a one-day to a one-month horizon. This, too, is consistent with micro-based models: At longer horizons, customer flows from a single bank should be a better proxy for the market-wide flow of information driving spot rates.

Micro-based models also make strong empirical predictions about the relationship between order flows and fundamentals. According to Eq. (1), dealers are forward-looking when quoting FX prices, so spot rates embody their forecasts for fundamentals based on common information, Ω_t^D. One empirical implication of this observation is that spot exchange rates should have forecasting power for fundamentals. While there is some evidence that this is true for variables that comprise fundamentals in many models (Engel and West 2005), the forecasting power is rather limited. Micro-based models also have implications for the forecasting power of order flows. If order flows convey information about fundamentals that is not yet common knowledge to all dealers (that is, not in Ω_t^D), then they should have incremental forecasting power for fundamentals, beyond the forecasting ability of any variable in Ω_t^D. This is a strong prediction: it says that order flow should add to the forecasting power of all other variables in Ω_t^D, including the history of spot rates and the fundamental variable itself. Nevertheless, Evans and Lyons (2004b) find ample support for this prediction using customer order flows and candidate fundamental variables such as output, inflation and money supplies. These findings provide direct evidence on the information content of customer order flows, and provide a new perspective on the link between exchange rates and fundamentals.

Dispersed information concerning fundamentals need not come only from the activities of individuals, firms and financial institutions. Scheduled announcements on macroeconomic variables (for example, GDP, inflation or unemployment) can also be a source of dispersed information. If agents have different views about the mapping from the announced variable to fundamentals, then the news contained in any announcement, while simultaneously observed, will not be common knowledge. For example, two firms may interpret the same announcement on last quarter's GDP as having different implications for future GDP growth. Differing interpretations about the implications of commonly observed news will be a source of customer order flows because they imply heterogeneous views about future returns, which in turn induces portfolio adjustment. Thus, micro-based models raise the possibility that the exchange rate effects of macro announcements operate via both a direct channel (that is, when the announcement contains CK news) and an indirect channel. Love and Payne (2003) and Evans and Lyons (2003, 2005b) find evidence that both channels are operable. Evans and Lyons estimate that roughly two-thirds of the effect of a macro announcement is transmitted indirectly to the dollar/deutschmark spot rate via order flow, and one-third directly into quotes. With both channels operating, macro news is estimated to account for more than one-third of the variance in daily depreciation rates. This level of explanatory power far surpasses that found in earlier research analysing the impact of macro news on exchange rates (for example, Andersen et al. 2003). It also further

cements the link between spot rates and the macro variables comprising fundamentals.

Order Flows, Returns and the Pace of Information Aggregation

The process by which the information contained in the customer flows becomes known across the market, and hence embedded into FX quotes, is complex. The individual customer and inter-dealer orders received by each dealer contain some dispersed information about the economy, but extracting the information from each order constitutes a difficult inference problem. Under some circumstances the inference problems are sufficiently simple for every dealer to learn all there is to know about fundamentals in a few rounds of inter-dealer trading. In this case, the pace of information aggregation is very fast, so that new information concerning fundamentals is quickly reflected in dealer quotes whether the news is initially dispersed or common knowledge. The resulting dynamics for exchange rates over weeks, months or quarters will be indistinguishable from the predictions of macro models. Under other circumstances, the inference problem facing individual dealers is sufficiently complex to slow down the pace of information aggregation. Here it takes many rounds of inter-dealer trading before the dispersed information concerning fundamentals becomes known across the market. This scenario is much more likely from a theoretical perspective. Evans and Lyons (2004a) show that the conditions needed for fast information aggregation are quite stringent. Of course, because inter-dealer trading takes places continuously, dispersed information could be completely embedded in FX quotes in a short period of calendar time (for example, a day), even if the pace of information aggregation is slow. In principle, dealers might be able to learn a good deal from the multitude of orders they receive in a typical day, even if individual orders are relatively uninformative. The question of whether it takes significant amounts of calendar time before dispersed information is embedded in FX quotes can be answered only empirically.

If the pace of information aggregation is slow, customer order flows across the market contain information that will become known to all dealers only at a later date. So, if the customer orders received by an individual bank are representative of the market-wide flows, they should have forecasting power for the future market-wide flow of information that drives quote revision. Recent empirical findings support this possibility. Evans and Lyons (2005b, c) show that customer order flows have significant forecasting power for future depreciation rates both in and out of sample. These results are qualitatively different from the contemporaneous empirical link between order flows and depreciation rates discussed above. In the context of Eqs. (2) and (3), the market-wide flow of information from period-t trading impacts on the depreciation rate, Δs_{t+1}, via ε_{t+1}. The contemporaneous link arises because period-t inter-dealer order flows measure the market-wide information flow, ε_{t+1}. In contrast, the forecasting power of customer flows for the depreciation rate arises because ε_{t+1} contains information that was originally in the customer orders received by individual banks *before* period-t trading.

These forecasting results are surprising in terms of both their horizon and strength. In particular, out-of-sample forecasts based on customer flows from month $t-1$ can account for roughly 16 per cent of the variation in next month's depreciation rate, Δs_{t+1}. This finding suggests that the pace of information aggregation is far, far slower than was previously thought; it seems to take weeks, not minutes, for dispersed information to be fully assimilated across the market. The level of forecasting power is also an order of magnitude above that usually found in exchange rate models. For example, the in-sample forecasting power of interest differentials for monthly depreciation rates is only in the two to four per cent range.

The slow pace of information aggregation may shed light on one of the long-standing puzzles in exchange rate economics; the disconnect between spot exchange rates and fundamentals over short and medium horizons (Meese and Rogoff 1983). The idea is quite simple. If changes in fundamentals are reflected in spot rates only when information concerning the change is recognized by

dealers across the market, the slow pace of information aggregation will mask the link between the depreciation rate and the change in fundamentals over short horizons, because the latter is a poor proxy for the market-wide flow of information. Simulations in Evans and Lyons (2004a) show that this masking effect can be quite substantial. Fundamentals account for only 50 per cent of variation in spot rates at the two-year horizon even though information aggregation takes at most four months.

One factor that might contribute to the slow pace of information aggregation is the presence of price-contingent order flow generated by feedback trading. Stop-loss orders, for example, represent a form of positive feedback trading in which a fall in the FX price triggers negative order flow from customers wishing to insure their portfolios against further losses. Feedback trading of a known form does not complicate the inference problem facing dealers because the orders it generates are simply a function of old market-wide information. However, when the exact form of the feedback is unknown it makes inferences less precise and so slows down the pace of information aggregation. Osler (2005) argues that feedback trading will be an important component of order flow when quotes approach the points at which stop-loss orders cluster. A fall in FX quotes at these points can trigger a self-reinforcing price cascade where causation runs from quotes to order flow.

Some economists argue that the early empirical findings linking order flow and the depreciation rate reflected the presence of positive feedback trading rather than the transmission of dispersed information. Indeed, there is no way to tell whether intra-day causation runs from order flows to quotes or vice versa from just the contemporaneous correlation between order flow and the deprecation rate measured in daily data. However, the new evidence on the forecasting power of order flow for both depreciation rates and fundamentals firmly points to order flow as the conveyor of dispersed information. This is not to say that feedback trading is absent. Portfolio insurance and other price-contingent trading strategies (such as liquidity provision) undoubtedly contribute to order flows, and their presence may actually explain why the pace of information aggregation is so slow.

Future Research

Exchange rate research based on micro-based models is still in its infancy. The past few years have seen a rapid advance in theoretical modelling and some surprising empirical results. Advances on the empirical side will be spurred by the greater availability of trading data. On the theoretical side, micro-based modelling may provide new insights into the determinants of the foreign-exchange risk premium, the efficacy of foreign exchange intervention, and the anatomy of financial contagion.

See Also

- ▶ Exchange Rate Dynamics
- ▶ Exchange Rate Volatility
- ▶ Information Aggregation and Prices

I thank Richard Lyons for valuable discussions and gratefully acknowledge the financial support of the National Science Foundation.

Bibliography

Andersen, T., T. Bollerslev, F. Diebold, and C. Vega. 2003. Micro effects of macro announcements: Real–time price discovery in foreign exchange. *American Economic Review* 93: 38–62.

Bacchetta, P., and E. van Wincoop. 2006. Can information heterogeneity explain the exchange rate determination puzzle? *American Economic Review* 96: 552–576.

Engel, C., and K. West. 2005. Exchange rates and fundamentals. *Journal of Political Economy* 113: 485–517.

Evans, M., and R. Lyons. 2002a. Order flow and exchange rate dynamics. *Journal of Political Economy* 110: 170–180.

Evans, M., and R. Lyons. 2002b. Informational integration and FX trading. *Journal of International Money and Finance* 21: 807–831.

Evans, M., and R. Lyons. 2003. How is macro news transmitted to exchange rates? Working Paper No. 9433. Cambridge, MA: NBER.

Evans, M., and R. Lyons. 2004a. A new micro model of exchange rates. Working Paper No. 10379. Cambridge, MA: NBER.
Evans, M., and R. Lyons. 2004b. Exchange rate fundamentals and order flow. Working paper. Online. Available at http://www.georgetown.edu/faculty/evansm1/. Accessed 16 Nov 2007.
Evans, M., and R. Lyons. 2005a. Do currency markets absorb news quickly? *Journal of International Money and Finance* 24: 197–217.
Evans, M., and R. Lyons. 2005b. Meese–Rogoff Redux: Micro-based exchange rate forecasting. *American Economic Review, Papers and Proceedings* 95: 405–414.
Evans, M., and R. Lyons. 2005c. Exchange rate fundamentals and order flow. Mimeo, Georgetown University. Online. Available at http://www.georgetown.edu/faculty/evansm1/. Accessed 6 June 2006.
Froot, K., and T. Ramadorai. 2005. Currency returns, intrinsic value, and institutional–investor flows. *Journal of Finance* 60: 1535–1565.
Hau, H., and H. Rey. 2006. Exchange rates, equity prices, and capital flows. *Review of Financial Studies* 19: 273–317.
Love, R., and R. Payne. 2003. Macroeconomic news, order flows, and exchange rates. Discussion Paper No. 475. Financial Markets Group, London School of Economics.
Lyons, R. 1997. A simultaneous trade model of the foreign exchange hot potato. *Journal of International Economics* 42: 275–298.
Meese, R., and K. Rogoff. 1983. Empirical exchange rate models of the seventies. *Journal of International Economics* 14: 3–24.
Osler, C. 2005. Stop-loss orders and price cascades in currency markets. *Journal of international Money and Finance* 24: 219–241.
Payne, R. 2003. Informed trade in spot foreign exchange markets: An empirical investigation. *Journal of International Economics* 61: 307–329.

Foreign Exchange Markets, History of

Marcello de Cecco

Abstract

Foreign exchange transactions, known in classical antiquity, developed into markets in the Middle Ages. Italian dealers dominated the market until the 16th century, when they started being replaced by the Dutch and English. The City of London has been the centre of world forex markets since the 18th century and remains dominant even today. Transaction modes have been revolutionized by information technology. Volume has also grown enormously. But personal contact is still important, hence financial centres persist. The arrival of the euro has had consequences for the forex market are discussed, as will the emergence of China.

Keywords

Arbitrage; Bills of exchange; China, economics in; Derivatives; Euro; European Central Bank; European Monetary Union; Fiat money; Foreign exchange controls; Foreign exchange markets; Information technology; Interest; Specie-flow mechanism; Speculation; Usury

JEL Classifications

F3

'Foreign exchange markets' is an expression that people normally associate with foreign currency transactions, whether in notes or coins. That association is correct, but foreign exchange markets trade in all transactions concerning debt instruments denominated in foreign currencies. This is not a modern development, even if it is true that debt instruments and the transactions associated with them have multiplied as economies have become more complex and more open to one another.

Origins and Causes

Trade in coins and debt instruments denominated in foreign currency is an ancient activity. Reference to it is found in ancient literatures and inscriptions belonging to many different cultures. From what one can glean from these ancient texts, it was always a type of trade organized by dealers, who were sometimes only brokers but more often than not traded for their own account, and often mixed foreign exchange dealing with merchandizing and lending.

In the development of foreign exchange activities, it is impossible to exaggerate the importance exercised by the Aristotelian prohibition of usury, which the Koran and scholastic doctrine

perpetuated in Muslim and Christian lands. Aristotle thought that only living beings could bear fruit. Money, not a living being, was by its nature barren, and any attempt to make it bear fruit (*tokos,* the Greek for 'bearing fruit', also means 'interest') was a crime against nature.

The need for intertemporal planning of economic activities requires the use of lending and its remuneration. Human ingenuity discovered, very soon after the Aristotelian prohibition, that while lending gave rise to interest (which was against nature), the sale of one asset against another, including coins, was a legitimate activity. Hence, the price at which that sale occurred could very appropriately hide a lending transaction. There followed an enormous diffusion of asset sales–purchases, which, after the break-up of the Roman Empire in the fifth century AD and the fragmentation of the Roman currency area into many smaller zones, often became foreign exchange transactions. The fluctuation of exchange rates between currencies provided a convincing case of risk associated with foreign exchange activities, and further reduced the possibility of transactors being accused of usury.

Raymond De Roover (1954) attributes to the Aristotelian prohibition the redirection of banking towards foreign exchange transactions that occurred from the early Middle Ages onwards. Since lending and borrowing at interest were outlawed, they had to be hidden inside more and more imaginatively devised foreign exchange transactions. This is a perfect case of financial innovation spurred by legal prohibition, which acquires a momentum of its own, generating a huge crop of by-products. Most of these by-products, and foreign exchange contracts and practices devised in the Middle Ages, are still present in today's markets, often even keeping their original names.

The most typical case is that of the bill of exchange, which is a transaction between two or more agents, giving rise to an exchange of foreign currency to be effected in different places at different times. The multiplicity of transactors, and of the contract's attributes, allows the fashioning of the contract in a remarkable number of different ways, following the needs of the transactors and the development of commercial and banking habits.

The fact that foreign exchange transactions are sales–purchases of assets denominated in foreign currencies, and that for a long time what could easily have been transacted in one currency had to be hidden behind a foreign exchange transaction, contributed from early on to the weaving of foreign exchange theory into an intricate web, as trade flows were recognized to be just one of the factors determining foreign exchange rates. Asset transactions obviously contributed at least as much to their determination. But while trade was visible, asset sales–purchases were not easily detected and recorded, and it was much more difficult to attribute exchange rate oscillations to their influence. This was especially so if, as we have already noted, a great number of such foreign exchange transactions, giving rise to a large volume of bills of exchange, actually hid domestic lending activity.

Foreign Exchange in the Middle Ages: The Rise of Italian Market Supremacy

The fragmentation of the Roman Empire gave rise in Italy to a fragmentation of monetary sovereignty and to the accompanying early specialization of Italian merchants in foreign exchange transactions. The fact that the papacy was also seated in Italy made the adherence to religious prohibitions of usury superficially stricter; but, with the help of scholastic doctors, merchants were able to devise ways to circumvent the prohibitions.

All this ended up in helping Italian merchants to develop a vast body of knowledge about foreign exchange banking, which they tried to keep to themselves for as long as they could. Thus they became specialists in the transfer of funds from one place to another. Difficulty and danger connected with travel discouraged the physical transportation of metallic money, which was a scarce good anyway, at least until the diffusion of fiat currency in the 19th and especially the 20th centuries. Whoever could transfer titles to assets between geographically distant places stood to gain a great deal of money and power. Italians became masters of these arcane practices. First Florentine and Venetian bankers, then the Genoese, practically cornered this market for

several centuries. They developed an enormous clearing network, encompassing most relevant trading places, where they kept agents and correspondents. As a result they could effect transfers everywhere. Sovereign rulers, who had to transfer vast sums because of their military operations in foreign lands, were the Italian bankers' best and worst clients. They tried to escape from the bankers' clutches, and to foster competition, but more often than not they were forced back into the bankers' hands by the superiority of the Italians' skills and by the bankers' monopoly of power. Philip II of Spain confided in a letter his dismay at not being able to understand foreign exchange problems. He had tried to get rid of the Genoese, but had to accept soon after that only they were able to transfer his American treasure from Spain to Flanders, and thus circumvent the maritime power of the English.

The Market Shift to Atlantic Europe

With the decline of the religious condemnation of usury, and the shift of trade from the Mediterranean to the Atlantic, the Italians' tight monopoly on the foreign exchange market faded away and was transferred first to Belgium and the Low Countries and then to Great Britain, or, more precisely, to London. It is quite remarkable how this skill always managed to bypass France, despite its being the richest country in Europe. Champagne fairs were dominated by foreign merchants, who monopolized exchange transactions. The same was true in Lyons. In fact, even the transfer of foreign exchange transactions to the shores of the North Sea and the Atlantic should be seen largely as a physical relocation of foreign exchange specialists to the places where trade had flourished. Foreign exchange transactions have remained a footloose activity, practised by a close-knit coterie of specialists who can move their show to where conditions are favourable, decamping without much ado from places where regulators have become too nosy or fiscal requests too oppressive. This is true even in this day of huge national banks and powerful central banks. It was even more apparent when those institutions were in their infancy and international bankers roamed the world free, holding sovereign rulers in their power.

The City of London's Market Supremacy

The monopoly that the Italians held over foreign exchange transactions was reproduced in more modern times by the City of London, where even today the largest concentration of such transactions takes place. British bankers have presided over most of the innovations that have taken place in this market because of the development of modern technologies. Everybody has heard of the homing pigeon informing the House of Rothschild of the outcome of Waterloo – such was the state of information transmission at the beginning of the 19th century. In the second half of that century, however, technical progress in this field advanced by leaps and bounds, revolutionizing foreign exchange technology. Distance between markets and the slow flow of information had meant that interest rate differentials between different financial markets could remain open for months before being noticed and closed by foreign financial flow. Arbitrage activity had thus been linked, more than to anything else, to seasonal patterns, as one easily discovers by reading contemporary treatises. It was noticed that money was recurrently scarce in one particular month or season in one specific market. Merchants would contribute to fill the gap, if enough profit was expected from transferring money from other places. Alternatively, or concurrently, the Humean specie-flow mechanism would intervene to transform this money scarcity into increased exports and imports. With faster flow of information made possible first by the steam engine, then by the telegraph, then by the international and intercontinental cable, and finally by the radio and telephone, the arbitrage margins between different financial markets came to be closed at speeds that could not be compared with earlier times. This became particularly apparent from the end of the 19th century. However, the vast increase in the speed and volume of foreign exchange transactions which has accompanied innovation in information technology appears to have given just as much chance to foreign exchange speculation, linking together asset markets that had

previously remained purely domestic, and by mixing speculation in foreign exchange with commodity speculation in a volume that could not have been attained in previous times.

Inception of Foreign Exchange Controls

Given the advances in information technology, the prevalence of speculation over arbitrage could have generated major international financial crises and so endangered the work of the international economy as much as the advances in information technology had enhanced it. The realization of these dangers, and the palpable loss of monetary sovereignty which the linking of financial markets brought in its train, convinced economic authorities in the period between the two world wars to try to isolate their respective national financial markets by foreign exchange restrictions. Although they were practised with great fervour and severity in Britain too, after the Second World War the City of London managed to persuade the authorities to get rid of them and give the City a chance to go back to its earlier domination of the commodities and exchange markets. In spite of the emergence of New York, Tokyo, and Frankfurt as prime financial markets, the hold British bankers have managed to keep over commodities and exchange transactions is indeed remarkable, and can be considered equal in length of time, breadth and intensity only to that previously exercised on the same activities by the Italian bankers.

Persistence of London's Supremacy

This persistence, in the face of the obvious decline of British and previously Italian economic power, is extremely interesting. The commodities and foreign exchange markets seem to have successfully ridden, and to have used to their benefit, the momentous advances in information technology which came in waves in the 19th and 20th centuries. It was expected for these advances to enhance the diffusion of such transactions, by de-concentrating and de-monopolizing them. This of course has happened, but not nearly to the extent that was expected. Technical innovations have also been used to reinforce market power. Having the dollar as a reserve currency has not seemed to help New York become the home of the commodities and foreign exchange markets. Nor does the demotion of sterling seem to have unduly penalized the City as far as those markets are concerned.

It is obvious that some of the reasons that have brought merchants to congregate in certain places ever since early times persist even in the age of global real-time transactions. Physical proximity and cultural affinity are still powerful enhancers of smooth and successful transactions, as is the confidence that the government will not disturb operations with crippling regulations or with oscillatory behaviour, which destroys certainty. It is perhaps this unique mix of factors that makes for the permanence of foreign exchange markets in certain places. Other pillars of economic power, like a great industrial structure, seem to be somewhat inimical to the permanence of commodities and foreign exchange markets in a given place. Industry certainly generates exports and foreign exchange transactions; but it soon also develops credit needs of its own, and possibly protectionism, both of which work against the permanence of a foreign exchange market. Governments are asked by industry to adopt policies that go against the total freedom that commodities and exchange markets require in order to thrive. Their adoption of such policies induces the community of foreign exchange dealers to pitch its tents elsewhere.

Market Growth Since the 1980s

This plea for continuity in history must not, however, be to the detriment of realism – and realism imposes a thorough appreciation of the huge increase the foreign exchange market has experienced since the 1980s. As we have already noticed, computer power increased prodigiously in the 1980s and permitted the real-time connection of forex markets across time zones, in a temporal and geographical continuum. Computer power also allowed ever more sophisticated forex contracts to be priced in real time and thus to be executed very rapidly. Among the more exotic

contracts, so-called derivatives must be mentioned, which further contributed to increasing the size of forex markets. The size of the market in 2007 is estimated by the Bank for International Settlements (BIS) to be around two trillion dollars.

As we noted above, continuity remains a feature of this huge market. London is still the place where more than 25 per cent of all transactions are processed, with New York coming a distant second, and Tokyo an even more distant third.

And, in spite of the huge size of the market, a few giant international banks concentrate a remarkable percentage of total transactions. The ten largest dealers account for 70 per cent of total transactions. Six of them are commercial banks and four are investment banks.

How the Market Looks Today

At the turn of the millennium, the euro was introduced, an important novel type of currency, not the expression of a sovereign state, but issued by the European Central Bank on behalf of the European Union. This innovation profoundly changed the forex market, as it marked the disappearance of all transactions denominated in the currencies of the European Monetary Union member states, and it meant the arrival of a dominant currency pair the euro/US dollar pair, which in 2004 already accounted, according to the BIS (2004), for 28 per cent of all forex transactions, followed by the US dollar/Japanese yen pair, which accounted for 18 per cent of all transactions. Remarkably, in 2004 14 per cent of all transactions were still taking place between the US dollar and the British pound. Since 1985 most British merchant banks have been swallowed up by foreign financial institutions, mostly commercial banks, but the British pound, and London, remain foreign exchange favourites.

The present situation in the forex market thus bears an important echo of past power, in the persistence of the British pound, a testimony of recent world economic events, with the arrival and very rapid establishment of the euro as dominant instrument for forex transactions, and of the Japanese yen as the third most important currency. Almost no trace is yet to be seen in the forex market of the meteoric rise of China on the world economic scene. The Chinese currency has recently gained some current account convertibility, but it will be years before it becomes fully convertible. Until then, it will not be able to form important currency pairs with the other dominant currencies. This should come as no surprise if we remember how many years it took the yen to establish itself in the position it now enjoys in the forex market. It should also constitute a final and conclusive piece of evidence in favour of what was noted above on the international foreign exchange community's susceptibility to national fetters and regulations.

See Also

- ▶ Foreign Exchange Market Microstructure
- ▶ Gold Standard

Bibliography

BIS (Bank for International Settlements). 2004. *Triennial Central Bank survey 2004*. Basel: BIS.

De Roover, R. 1954. New interpretations of the history of banking. *Journal of World History* 4: 38–76.

Kindleberger, C.P. 1984. *A financial history of western Europe*. London: Allen & Unwin.

Mandich, G. 1953. *Le Pacte de Ricorsa et le Marché Italien de Changes un XVIIe Siècle*. Paris: Armand Colin.

Foreign Exchange Reserve Management

Claudio Borio

Abstract

Foreign exchange reserve management refers narrowly to the allocation of foreign exchange

The views expressed are those of the author and do not necessarily reflect those of the BIS.

reserves across currencies, asset classes and instruments. Reserve management practices have evolved substantially over the first decade of the twenty-first century, with a tendency for processes to converge to those in the private asset management industry. There has been a greater tendency to focus on return, a more structured allocation process, and a greater focus on risk management, including reputational as well as pure financial risks. Decisions on the currencies in which to hold reserves and to designate as numeraire (currency of account) are also receiving increasing attention.

Keywords

Currency; Foreign exchange; Reserve management; Portfolio management; Numeraire

JEL Classifications

E58; F30; F31; G11; G15

The expression 'foreign exchange reserve management', strictly defined, refers narrowly to the allocation of foreign exchange reserves across currencies, asset classes and instruments. Thus, it excludes decisions concerning the level of reserves and foreign exchange intervention. And since some of the reserves may be 'borrowed', the stricter definition takes as given the net foreign exchange position (exposure to foreign exchange risk). Institutionally, some ambiguity exists as a result of the growth of publicly owned funds specializing in investments in foreign currency assets (so-called sovereign wealth funds, SWFs). These funds are sometimes invested in liquid assets and established partly as carve-outs of portfolios officially defined as foreign reserves. For present purposes, however, SWFs are excluded from the analysis.

Historically, economists have been much more interested in the question of what determines the overall level and rate of change of foreign exchange reserves than in what determines their composition. It is the level and the rate of change that are more closely related to the ability of a country to insulate itself from external shocks, to its wherewithal to provide lender-of-last-resort support to domestic financial institutions in need of foreign currency, and, through exchange market intervention and the corresponding sterilization decisions, to the determination of exchange rates and the level of domestic interest rates. Issues related to the composition have remained mainly the preserve of those in charge of reserve management policies, mostly central banks, and of market practitioners. Even so, the marked acceleration in the growth of world reserves since the mid-1990s and the emergence of some very large players, not least in Asia, have greatly added to the overall interest. In some cases, old questions have gained new salience: could the US dollar lose its status as the unrivalled reserve currency, and lose value in the process, owing to reserve diversification? In other cases, questions were asked for the first time: could allocation decisions have a first-order effect on asset prices, such as by pushing down long-term yields as a result of diversification away from short maturities? How far should countries seek higher-yielding returns to limit the opportunity cost of holding such large stocks of reserves?

Reserve management practices have evolved substantially over the first decade of the twenty-first century, to the point that some of the older economic analysis of the subject has become rather uninformative and potentially misleading. The overarching trend has been a tendency for reserve management processes to converge to those in the private asset management industry. This trend has manifested itself in at least three ways.

First, there has been a gradual shift towards more return-oriented strategies. Liquidity and safety (capital preservation) have traditionally been the primary objectives of reserve managers. Over time, however, greater attention has been paid to raising returns. Reserve managers have gradually broadened the range of asset classes they can invest in, and have shifted the portfolio composition towards higher risk allocations. For example, bank deposits and treasury bills have partly given way to longer-term bonds and agency paper; and some portfolios have been broadened to include also asset-backed securities, corporate debt and even equities. In the process, gold holdings have lost

ground. The management of this wider investment universe has led to greater reliance on external managers. There has also been a growing tendency to divide the reserves portfolio into tranches managed with different objectives, with some focused more on liquidity and others on investment returns. Reserve managers have increasingly been making use of derivatives. And in some cases, in order to gain greater room for manoeuvre, reserves have been transferred to SWFs, where the funds can be managed less conservatively.

Second, reserve management has been implemented through a much more structured decision-making process. First and foremost, there has been a move towards a more top-down approach. In the past, the overall asset allocation tended to result from the passive aggregation of decisions of individual traders, subject to strict limits. Over time, the executive level has become more directly involved in defining the acceptable risk–return trade-off for the reserve portfolio (the strategic asset allocation, SAA), articulated by selecting a 'benchmark' portfolio and by defining the tolerance ranges within which the actual allocation is allowed to vary. In addition, a growing number of central banks have been putting in place an intermediate layer between the SAA level and portfolio management execution level, at which decisions aimed at exploiting shorter-term market developments are taken: a tactical asset allocation (TAA) level. Alongside the more top-down approach, there has been a tendency to increase the functional separation of the activities involved in the reserve management process ('horizontal separation'). The objective has been to strengthen its integrity, by limiting opportunities for actual or perceived conflicts of interest across trading, performance measurement, risk analysis and settlement functions.

Finally, risk management has been strengthened. In the design of the portfolio, in particular the SAA, more rigorous analysis of market risks over the relevant horizons has become increasingly important. Subject to the inevitable estimation/calibration problems, the role of quantitative analysis has risen. In the implementation of the portfolio, there has been a tendency to track more closely actual exposures relative to the desired targets and to the permissible tolerance ranges. The shift has been supported by increasingly sophisticated tools, such as stress tests/scenario analysis, Value-at-Risk (VaR) and tracking error analysis, used to measure the overall risk in the portfolio or the volatility around target allocations. Likewise, the measurement and management of credit and operational risks, historically more important than in the private sector because of the public authorities' high sensitivity to reputation risk, has also been upgraded along similar lines.

Several factors have supported these trends. For some countries, the unprecedented accumulation of reserves has been important. It has naturally encouraged a shift towards more return-oriented strategies and added to pressures for greater accountability in the management of a growing fraction of a country's resources. More generally, the development of financial markets and financial technology has improved the trade-off between liquidity and return while providing the tools for more rigorous asset allocation and risk management. Likewise, the broad trend towards greater central bank independence has increased the emphasis on accountability and transparency, encouraging a strengthening of both internal and external governance, including through greater disclosure. Moreover, the upgrading of internal governance and the adoption of more return-oriented strategies have been mutually supportive. Thus, while the degree of risk tolerance may naturally wax and wane with economic conditions and actual loss experience, the emphasis on a more structured decision-making process and on risk management is unlikely to be reversed.

Within this broad picture, some perspective and differentiation are needed. For one, despite the shift towards higher returns, foreign exchange reserve allocations remain quite conservative. A few currencies dominate allocations, and changes over time in these allocations have been comparatively small. Similarly, asset classes yielding a long-term risk premium generally account for only a small share of aggregate portfolios. In addition, despite greater convergence, there is considerable differentiation across countries in terms of the degree of risk tolerance and the structure of decision making.

Moreover, for all the growing similarities with the management of private sector portfolios, the criteria underlying foreign reserve holdings are, or at least should be, substantially different. The differences largely derive from the purposes for which reserves are held and their relationship to the other public sector functions performed by the reserve managers, mostly central banks. These differences are reflected in the difficulties faced in defining a proper risk–return trade-off for the portfolio and the related choice of the appropriate 'numeraire' currency to measure risks and returns.

The apparent 'conservative bias' in the management of foreign exchange reserves, for example, derives directly from the fact that the overriding goals of central banks, typically couched in terms of monetary, financial and macroeconomic stability, impose serious constraints on reserve management operations. Losses on the foreign exchange portfolio matter not so much for their intrinsic pecuniary value. Rather, the main concern is with their impact on either the institution's reputation – such as any risk of a charge of incompetence – or on other factors that might undermine its operational effectiveness, such as by threatening budgetary independence from the government. Similarly, financial gains that may be obtained at the expense of potentially destabilizing markets in other jurisdictions, such as by selling assets in a falling market, or by investing in asset classes denounced as too risky and as a threat to financial stability, could undermine the central bank's reputation. More generally, the true economic return on the reserves bears only a weak relationship to their financial return. It should ultimately be measured in terms of improved economic performance of the economy as a whole.

The choice of numeraire currency (unit of account) plays a key role in any asset management decision, since the numeraire is the unit in which returns and risks are measured. In order to limit risks, allocations will be heavily tilted towards currencies that are comparatively stable in relation to the numeraire. Taken for granted in private asset management (the 'domestic currency'), the choice of numeraire for reserve management is not straightforward. Ultimately, it should be determined based on the ultimate uses of the reserves (for instance: foreign exchange intervention, insurance of foreign goods and services, hedging capital account transactions and insulation from financial crises) or the consequences of the holdings for the central bank (for example, losses that may undermine its independence). Different considerations point to different choices, ranging from the most liquid foreign currency, to foreign currency baskets and to the domestic currency. The final choice will also have implications for the importance of the choice of foreign exchange regime in the currency allocation. For example, if the domestic currency is used as numeraire – as appears to be increasingly the case – the allocation will be heavily tilted towards the foreign currency, or basket, with respect to which the domestic currency is most stable.

The conceptual challenges faced in addressing the two issues just outlined are just one example of the many questions still outstanding in foreign exchange reserve management. For instance, how should the relevant portfolio be defined, and how far should its management take into account aspects of the private sector balance sheet? How far should the management of reserves be integrated with the rest of the central bank balance sheet? In fact, should foreign exchange reserves be managed as part of the broader public sector balance sheet or on their own? What are the most appropriate governance arrangements and degree of disclosure? While the basic analytics of these questions are common, answers to them are likely to be countryspecific. They are also bound to continue to evolve in light of changing intellectual paradigms as well as economic and political circumstances.

See Also

- ▶ Exchange Rate Dynamics
- ▶ Exchange Rate Exposure

Bibliography

Bakker, A.F.P., and I.R.Y. van Herpt. 2007. *Central bank reserve management: New trends, from liquidity to return*. Cheltenham: Edward Elgar.

Ben-Bassat, A. 1984. *Reserve-currency diversification and the substitution account*, Princeton Studies in

International Finance no. 53. Princeton: Princeton University, International Finance Section.
Bernadell, C., J. Coche, F.X. Diebold, and S. Manganelli. 2004. *Risk management for central bank foreign reserves*. European Central Bank.
Borio, C., J. Ebbesen, G. Galati, and A. Heath. 2008a. FX reserve management: Elements of a framework. BIS Papers no. 38, March.
Borio, C., G. Galati, and A. Heath. 2008b. FX reserve management: Trends and challenges. BIS Papers no. 40, May.
Dooley, M., S. Lizondo, and D. Mathieson. 1989. The currency composition of foreign exchange reserves. IMF Staff Papers, June.
Eichengreen, B., and D.J. Mathieson. 2000. The currency composition of international reserves: Retrospect and prospect. IMF Working Paper 00/131, 1 July.
Galati, G., and P.D. Wooldridge. 2006. The euro as a reserve currency: A challenge to the pre-eminence of the US dollar? BIS Working Papers no. 218, October (forthcoming in *International Journal of Financial Economics*).
Heller, H., and M.D.K. Knight. 1978. *Reserve-currency preferences of central banks*, Essays in International Finance no. 131. Princeton: Princeton University, International Finance Section.
Pringle, R., and N. Carver. 2008. *How Central Banks manage reserve assets*. London: Central Banking Publications.
Ramaswamy, S. 2008. Managing international reserves: How does diversification affect financial costs? *BIS Quarterly Review* 45–56.
Rodrik, D. 2006. The social cost of foreign exchange reserves. *International Economic Journal* 20(3): 253–266.
Truman, E.M., and A. Wong. 2006. *The case for an international reserve diversification standard*, Working Paper 06–2. Washington, DC: Institute for International Economics.
Wooldridge, P.D. 2006. The changing composition of official reserves. *BIS Quarterly Review* 25–28.

Foreign Investment

Herbert G. Grubel

Defined narrowly, foreign investment is the act of acquiring assets outside one's home country. These assets may be financial, such as bonds, bank deposits and equity shares or they may be so-called direct investment and involve the ownership of means of production such as factories and land. Direct investment is considered to take place also if the ownership of equity shares provides control over the operation of a firm. Johnson (1970) has suggested the expansion of the concept of foreign investment so that it parallels the modern Fisherian approach and distinguishes physical, human and knowledge capital. Accordingly, schooling abroad and technology transfers through the purchase of patents and licences represent foreign investment broadly defined.

In the 19th century, foreign investment involved mostly the ownership of financial assets (Iversen 1936). After World War II direct foreign investment began to dominate and attract much theoretical and empirical research efforts of economists and the concerns of politicians (Hymer 1976; MacDougall 1960; Reddaway 1968a, b; Kindleberger 1968; Johnson 1970; Caves 1971; Dunning 1981; Vernon 1966). The brain drain, international technology transfers and international bank-lending occupied many researchers after the 1960s.

Motives for Foreign Investment

The most fundamental motive for foreign investment is the desire of wealth-holders to maximize the value of their portfolio or net worth. However, this basic motive has been clarified and extended by the inclusion of risk, and analysts now often consider risk-adjusted rated of return to wealth-portfolios as the main motive for foreign investment. Under this approach, foreign investment is possible even if the yield on assets abroad is expected to be lower than that on domestic assets simply because an imperfect correlation of changes in foreign and domestic yields is expected to increase the risk-adjusted rate of return to the entire portfolio (Grubel 1968). Numerous studies have documented the benefits from the international diversification of portfolios as well as direct investment holdings (Rugman 1979).

Direct Foreign Investment

There are other motives for the purchase of assets abroad. They involve either externalities or

market imperfections, which are internalized or eliminated by the multinational enterprise.

Technological externalities arise, for example, from the very high fixed costs in capital-intensive industries. In such industries great efficiency gains can be had by measures which stabilize operations at a high level of output. The ownership or control over suppliers and marketing permits firms in these industries to achieve such stabilization objectives which would be unattainable if separate owners pursued independent profit-maximization strategies. Given that raw materials, energy sources and finished-product markets often are located in different countries, vertically integrated companies in these industries frequently are multinational (Kindleberger 1968; Caves 1971).

Imperfections in factor-input markets which give rise to direct foreign investment are due to economies of scale, mainly those arising from the use of knowledge. Such knowledge is especially important in the design, production and marketing of differentiated consumer goods but many also involve management systems and information about customers and sellers. In addition, firms are motivated to own foreign production facilities in order to assure control over the quality of products and the maintenance of commercial secrecy. Furthermore, through direct foreign investment, firms are able to capture the international spillover effects of advertising expenditures.

The final major explanation of direct foreign investment involves distortions introduced by government policies. Tariffs and other protective devices as well as subsidies and taxes can create conditions under which it is more profitable to produce in, rather than export to, a foreign country.

The theory of direct foreign investment has been enriched by the analysis of additional, somewhat less-central issues. These involve the firms' choice of location, the decision to license rather than exploit technological assets through direct foreign investment, the legal forms of foreign ownership and the role of diversification. The usefulness of direct foreign investment as a method for diversification has been questioned in arguments which point to the opportunities of individual stockholders to obtain all the benefits of international diversification in their own portfolios, much like the Miller–Modigliani model questioned the need of individual firms to concern themselves with their capital structure. Some of the most useful insights about the nature of direct foreign investment have been gained by the analysis of reasons for its postwar growth (Kindleberger 1968).

Attempts have been made to capture most of the motives noted above under the concept of 'internalization' and the 'eclectic theory of direct foreign investment' (Dunning 1977). These approaches to the explanation of direct foreign investment have not been accepted widely, probably because the phenomenon is too complex to be captured adequately by the theory of internalization. The eclectic theory, on the other hand, is too broad by its inclusion of all of the many driving forces behind foreign direct investment (Black and Dunning 1982; Buckley and Casson 1976; Kojima 1978 – for reviews and marginal extension).

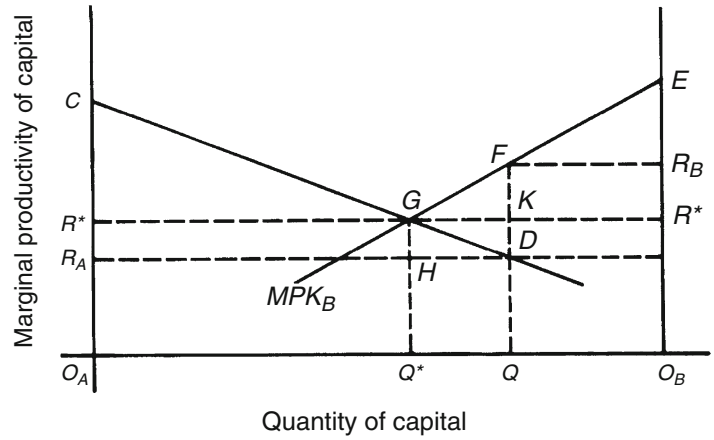
Empirical studies have found support for all of the motives noted above. While none dominates the others clearly, of some special importance appear to be economies of scale due to the ownership of knowledge capital and motives created by government.

Welfare Effects

During the 1960s concern over the welfare effects of foreign investment centred on its influence on the balance of payments as both the United States and the United Kingdom suffered from large and growing deficits. Two landmark studies (Reddaway 1968a, b; Hufbauer and Adler 1968) did much to sort out the different influences and interdependencies and produced some empirical estimates. Interest in the balance of payments effects of direct foreign investment has disappeared almost totally since the increased flexibility of exchange rates in the early 1970s.

Interest remains strong in the more general welfare effects of international investment, which received an influential early treatment by

Foreign Investment, Fig. 1



MacDougall (1960). The knowledge in this field can most easily be discussed with the help of Fig. 1, where the marginal productivity and the quantity of capital are on the vertical and horizontal axes, respectively. We assume that the world capital stock consists of the quantity O_A–O_B, and that there are two countries A and B, the marginal productivity of capital schedules of which are shown originating on the left and right side of Fig. 1 respectively. In the initial equilibrium before the opening of capital flows O_B–Q capital is held in country A and has a yield of O_A–R_A. Country B holds the rest of the capital with a yield of O_B–R_B.

Now assume that capital flows are permitted between the two countries and that as a result the owners of capital in country A invest Q^*–Q in country B. These investments reduce output in A and decrease it in B by the amounts Q^*QDC and Q^*QFG, respectively. As a result, rates of return are equalized in the two countries at Q^*R^*. Most important, the total productivity of the world's capital stock is increased by the area GDF. Such an output gain is the result of all capital movements, regardless of whether they take the form of investment in bonds, common shares, land, factories, human or knowledge capital.

In the new equilibrium the amount Q^*QKG represents the capital yield which accrues to its owners in country A, who therefore enjoy a net gain equal to the triangle GDK. Residents in the host country receive a net gain of GKF. Within country A the lowered capital–labour ratio raises the relative yield on capital and lowers that on labour. In country B the opposite effects take place.

As MacDougall (1960) pointed out, empirically the most important welfare effect of international capital-flows probably arises from taxation of profits and dividends by the country hosting the capital in combination with double taxation agreements which permit the foreign-tax payments to be deducted fully from tax obligations at home. In terms of Fig. 1, one half of the area Q^*QDH accrues to the residents of country B at the expense of the residents of country A, under the assumption that the tax rate is 50 per cent. This net gain to the host country is reduced by any subsidies or free services which its government provides to the foreign investment. Empirical studies of this taxation have shown it to involve large welfare effects (Grubel 1974).

Direct foreign investment often embodies new technology which cannot be acquired separately, and it leads to the net creation of workers' skills. In terms of Fig. 1 these effects result in an upward shift of the marginal productivity schedule of investment in the host country B. An additional area of output is created thereby, which accrues to the residents of the host country. Direct foreign investment can lead to increased competition in the host country and, through it, increased efficiency in the use of all domestic resources. Other,

more secondary welfare effects arise from changes in the two countries' terms of trade, which could go either way.

In models where there is unemployment equilibrium due to wage and price rigidities or underemployment as in developing countries, direct foreign investment can influence these conditions in both host and recipient countries. The efficiency models typified by Fig. 1 cannot deal with these welfare effects, and they have been relatively neglected in the literature under discussion here, even though they are of great concern to politicians.

Some Welfare Costs

The preceding neoclassical model has very little analysis of costs of direct foreign investment except for the tax effects and the usually peripheral issue of dynamic adjustment costs. This model has been attacked for neglecting several important ways in which direct foreign investment can reduce the welfare of the recipient country. Thus the owners of direct foreign investment can make investment, employment and output decisions that maximize rates of return but do not necessarily serve the interests of the host country; they can frustrate the achievement of monetary control as they draw on global capital sources; they can use their large resources to influence public opinion and elections in the interest of a foreign power or ideology; they compete unfairly with domestic producers who do not have access to low-cost capital and technology; they destroy domestic culture and traditions by the introduction of new and cheap goods, entertainment and art; they exploit monopoly and monopsony positions and thus charge too much for their products and pay too little for local inputs; they use transfer-pricing tricks to avoid the payment of host-country taxes; they create dependency on foreign supplies.

The evaluation of the preceding and many other arguments against direct foreign investment is difficult. Many of them are based on analytical paradigms which differ fundamentally from neoclassical economics. Others are based on empirical propositions that are nearly impossible to evaluate with available data. Still others involve value-judgements and implicit views on the relative efficiency of government substitute policies.

In the publications on foreign investment there is little interplay between the standard neoclassical approach to welfare effects and the analysis which stresses the costs. The former is taught and tends to dominate attitudes in industrial countries, while the latter is most popular and often very influential in developing countries and international organizations (Myrdal 1956; Hymer 1976; Behrman and Fischer 1980; Lall and Streeten 1977; United Nations 1973, 1978).

Policy Implications

The central policy issue in the field of international investment is whether or not it should be free, directed to achieve certain policy objectives or prohibited completely. The neoclassical paradigm implies that it should be free and that undesirable consequences accompanying it should be dealt with through policies directed at the problems themselves. As Bhagwati (1971) has shown, this approach permits the correction of market-failures without any sacrifice of the benefits from free trade in assets.

Other paradigms imply controls over foreign investment. At one extreme has been the complete prohibition of foreign investment in the Soviet Union and China after the Communist revolutions. These policies have been abandoned. Most countries of the world have some restrictions on foreign investment. Many insist that foreign investment has to be approved by a government agency, which uses acceptance criteria consistent with political and economic concerns of the time and the ruling party. Some countries restrict foreign ownership to minority holdings, which leave effective control with native entrepreneurs or governments. All of these restrictions involve costs of administration and diminish the level of international capital-flows. Therefore, they reduce the potential welfare gains below those attainable under the policy of dealing with market-failures directly.

See Also

- ▶ International Indebtedness
- ▶ Periphery
- ▶ Vent for Surplus

Bibliography

Behrman, J.H., and W.A. Fischer. 1980. *Oversees RD activity of transnational companies*. Cambridge, MA: Oelschlager, Gunn and Hain.

Bhagwati, J. 1971. The general theory of distortions and welfare. In *Trade, balance of payments and growth*, ed. J. Bhagwati et al. Amsterdam: Elsevier, North-Holland.

Black, J., and J.H. Dunning (eds.). 1982. *International capital movements*. London: Macmillan.

Buckley, P.J., and M.C. Casson. 1976. *The future of multinational enterprise*. London: Macmillan.

Caves, R.E. 1971. International corporations: The industrial economics of foreign investment. *Economica* 38: 1–27.

Caves, R.E. 1982. *Multinational enterprise and economic analysis*. Cambridge/New York: Cambridge University Press.

Dunning, J.H. 1977. Trade, location and economic activity and the multinational enterprise: A search for an eclectic approach. In *The international allocation of economic activity*, ed. B.P. Ohlin, P.O. Hesselborn, and P.J. Wiskman. London: Macmillan.

Dunning, J.H. 1981. Explaining the international direct investment position of countries: Towards a dynamic or development approach. *Weltwirtschaftliches Archiv* 117(1): 30–64.

Grubel, H.G. 1968. Internationally diversified portfolios: Welfare gains and capital flows. *American Economic Review* 58(5): 1299–1314.

Grubel, H.G. 1974. Taxation and rates of return from some US asset holdings abroad. *Journal of Political Economy* 82(3): 469–487.

Hufbauer, G., and M. Adler. 1968. *Overseas manufacturing investment and the US balance of payments*. Washington, DC: US Treasury Department.

Hymer, S.H. 1976. *The international operations of national firms: Study of direct foreign investment*. Cambridge, MA: MIT Press. (PhD dissertation, MIT, 1960).

Iversen, C. 1936. *International capital movements*. Oxford: Oxford University Press.

Johnson, H.G. 1970. The efficiency and welfare implications of the international corporation. In *The international corporation*, ed. C.P. Kindleberger. Cambridge, MA: MIT Press.

Kindleberger, C.P. 1968. *American business abroad*. New Haven: Yale University Press.

Kojima, K. 1978. *Direct foreign investment*. London: Croom Helm.

Lall, S., and P. Streeten. 1977. *Foreign investment, transnationals and developing countries*. London: Macmillan.

MacDougall, G.D.A. 1960. The benefits and costs of private investment from abroad: A theoretical approach. *Economic Record* 36(73): 13–35.

Myrdal, G. 1956. *Development and underdevelopment*. National Bank of Egypt: Cairo. Reprinted in *Leading issues in economic development*, ed. G. Meier. New York: Oxford University Press, 1970.

Reddaway, W.B. 1968a. *Effects of UK direct investment overseas*. Cambridge: Cambridge University Press.

Reddaway, W.B. 1968b. *Effects of UK direct investment overseas: Final report*. Cambridge: Cambridge University Press.

Rugman, A.M. 1979. *International diversification and the multinational enterprise*. Lexington: D.C. Heath.

Stopford, J.M., and J.H. Dunning. 1983. *Multinationals: Company performance and global trends*. London: Macmillan.

United Nations. 1973. *Multinational corporations in world development*. New York: United Nations.

United Nations. 1978. *Transnational corporations in world developments: A reexamination*. New York: United Nations.

Foreign Trade

Ian Steedman

The pure theory of trade constitutes, in principle, no more than an application of the general theory of value, distribution and resource allocation. It follows at once, of course, both that each possible approach to general economic theory has its corresponding theory of trade and that any changes or developments in general theory must have implications for the theory of international trade. In particular, this is true of certain debates over value, distribution and capital goods which flourished in the 1960s, following the publication of Piero Sraffa's *Production of Commodities by Means of Commodities* (1960).

It need hardly be said that capital goods – that is, produced inputs, whether they be long-lived or short-lived – are of the very greatest importance in all modern economies. And it is no less true that international trade flows, far from consisting

solely of consumption commodities, contain a large and growing volume of producer goods. International trade statistics are not conveniently classified into 'finished consumer goods' and 'other goods' but it appears from the classifications that are available that finished consumer goods probably account for less than some 30 per cent of the value of world trade. Any adequate theory of trade and resource allocation must, then, be able to deal, in a clear and coherent manner, with the important role of produced inputs and it is therefore to be expected that produced inputs would feature prominently in trade theory and that 'capital theory', broadly interpreted, should have significant implications for trade theory. But in fact, when we turn to basic trade theory, we find that capital goods are noticeable only by their absence, all the attention being centred on final consumption commodities.

With respect to capital theory, it is now well known that, in a competitive, constant-\returns-to-scale economy using produced inputs (a) relative prices depend on the rate of interest, even for a given technique; (b) capital-intensity depends on the rate of interest, even for a given technique; (c) the choice of technique need not be monotonically related to the rate of interest; and (d) capital-intensity, in a multi-technique economy, need not be inversely related to the rate of interest. (See, for example, the *QJE* Symposium 1966; Pasinetti 1977.) Also well-known are the results that, in an economy experiencing steady growth, there is a 'consumption-growth rate' trade-off which is identical to the 'wage-profit rate' frontier and that only if the growth rate equals the profit rate – the so-called Golden Rule case – is it ensured that the competitive choice of technique will be optimal with respect to the consumption/growth trade-off.

Suppose now that production is carried out using inputs of homogeneous land, as well as homogeneous labour, and produced inputs. Let there be a given, positive rate of interest on the value of capital (the produced inputs); it is then no longer the case that a rising rent/wage ratio must necessarily be associated with a falling land/labour ratio; quite the opposite relationship may hold (Metcalfe and Steedman 1972; Montet 1979). It follows that, in the presence of a positive rate of interest, an *increase* in the relative price of the more land-intensive commodity may be associated with a *decrease* in the output of that commodity (and an increase in the output of the labour-intensive commodity). In other words, there may be a 'perverse' supply response.

In brief, then, capital theory discussions have alerted us (or realerted us, for Wicksell (1901) was well aware of some of these complications) to the distribution-relative nature of relative commodity prices, to the fact that capital-intensity depends on distribution as well as on technical conditions, to the possibility that both capital-intensities and land-labour ratios may respond in 'unexpected' ways to changes in interest, wage and rent rates, to the fact that supply responses can differ from those traditionally supposed and to the possibility that competitive technique choice need not be optimal with respect to the consumption-growth rate trade-off. We now turn to the implications of these findings for the pure theory of trade.

'Textbook' Ricardian Theory

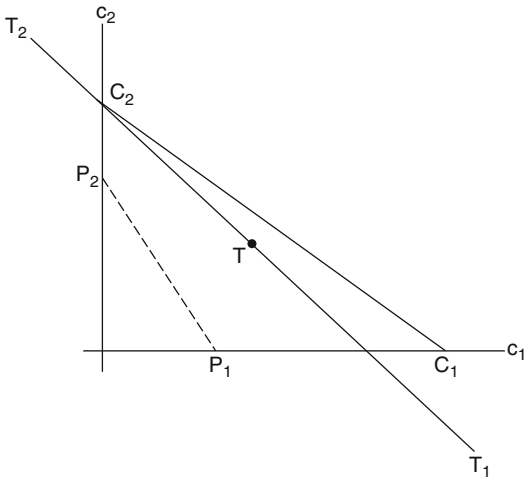
The reader will be thoroughly familiar with the textbook version of Ricardian trade theory, in which wages are the only kind of income, labour is homogeneous and – as a result of these two assumptions – the autarky price ratios in an economy are exactly proportional to the quantities of labour required to produce the various commodities. Yet when we turn to Ricardo's famous Chapter VII, 'On Foreign Trade' (1817), we see at once that Ricardo supposes there to be a *positive* rate of profit and, indeed, shows how the opening of trade can increase that rate. To this extent, then, 'textbook' Ricardian trade theory is a travesty of Ricardo's theory. Any attempt to excuse this vulgarization of Ricardo would probably appeal to the fact – and it is a fact – that in his Chapter VII Ricardo, whilst acknowledging the presence of both wages and profits, took no account of the influence of distribution on autarky relative prices; he simply identified these latter with relative labour quantities. Yet a large part of Ricardo's Chapter I, 'On Value', is concerned

precisely with the fact that, as was noted above, relative commodity prices depend on distribution and not on technical conditions of production alone. The apparent inconsistency is explained by Ricardo's readiness to assume that relative labour costs provide a 'good enough approximation' to relative prices, even though he fully acknowledged that prices really depend on distribution. This explanation, though, is not a justification of Ricardo's procedure in Chapter VII, for he gave quite inadequate grounds for his claim about the 'good enough approximation'. It follows that we should examine carefully what happens to Ricardo's propositions concerning foreign trade when full recognition is given to the distribution-relative nature of autarky prices.

Consider then a two-country, two-consumption commodity model in which, in each country, the autarky price ratio of the two consumption commodities *depends on* the ruling (r, w) under autarky. Such a dependence could arise from the use of (nontradeable) machines in making the consumption commodities; or from the fact that the consumption commodities are *also* capital goods, being used in the production of one another; or from the fact that wages are paid in advance and that the production period over which they have to be advanced differs as between the two consumption commodities. There are many different models which capture the dependence of relative prices on (r, w), all of them providing examples of what Samuelson (1975) has called 'time-phased Ricardian systems'. Now if, in either economy, the autarky rate of interest should happen to be zero, the autarky price ratio of the two consumption commodities will indeed equal the ratio of their total (direct and indirect) labour costs. This must be true when the only form of income payment is that of wages paid to homogeneous labour. But if, as will generally be the case, the autarky interest rate is not zero and fluke technical conditions do not obtain, that autarky price ratio will *not* equal the corresponding labour cost ratio.

Let free trade be opened between our two economies. Will the direction of trade be determined by a comparison of the two countries' autarky price ratios or by a comparison of their labour cost ratios? By the former, of course, since competition works via wages, interest rates and prices. Each country will export that commodity for which it has the lower relative autarky *price*. It may or may not export that commodity for which it has the lower relative labour cost and certainly the pattern of trade is not determined by technical conditions alone but depends also on the autarky (r, w) in each country, simply because autarky relative prices so depend. Notice the corollary that two economies with the *same* technical conditions, for producing commodities by means of homogeneous labour and produced commodities, could enter into free trade if their autarky (r, w) would be different. It is not the case that 'Ricardian' trade models must necessarily suppose different technical conditions in each country – even if it is the case both that Ricardo did make such an assumption and that it is eminently sensible to do so.

Consider now a single, small economy of the kind considered above, which faces given terms of trade for trade in the two consumption commodities. Its pattern of trade will depend on how the given terms of trade compare with its autarky *price ratio*. But whether its fully-specialized, free trade consumption bundle lies outside its autarky consumption-possibility-frontier will depend on that pattern of trade and on how the terms of trade compare with the economy's *labour cost ratio*. Since this latter ratio is not equal, in general, to the autarky price ratio, it is *not* ensured that the with-trade bundle will lie outside the autarky frontier. Consider Fig. 1, in which c_1 and c_2 are quantities of the first and second consumption commodities per unit of employment. C_2C_1 is the autarky consumption-possibility-frontier, whose absolute slope is of course equal to the labour cost ratio for the two consumption commodities. P_2P_1 is a line whose absolute slope is equal to the economy's autarky price ratio and T_2T_1 a line with slope equal to the given terms of trade. Since T_2T_1 is less steep than P_2P_1 the economy will be driven to specialize in commodity 2 – but, since T_2T_1 is steeper than C_2C_1, the economy's free trade consumption bundle, T, which must of course lie on T_2T_1, will be *below* the autarky frontier C_2C_1 (unless at C_2 itself). It will be clear that this result would not obtain if T_2T_1 were either steeper than



Foreign Trade, Fig. 1

P_2P_1 (with specialization at C_1) or less steep than C_2C_1 (with specialization at C_2). But the fact remains that Ricardo was able to be 'sure' about the gain from trade only because he illegitimately supposed C_2C_1 and P_2P_1 to have the same slope. This argument can be extended to a steadily growing economy, to show that in the 'Golden Rule' case the with-trade bundle must lie outside the achievable autarky frontier but that if the growth-rate is less than the profit-rate then it may or may not do so (as in Fig. 1, which provides simply a special case of this result, with a growth-rate of zero). Since the adoption of a particular specialization can, from a formal point of view, be thought of as a particular choice of technique, the present argument is just an application, to the trade context, of the capital theory result concerning competitive choice of technique and its possible non-optimality in terms of consumption and growth. It is important to notice that this result, concerning the possible (not certain) 'loss from trade', belongs to the class of 'comparative dynamics' results; it is best thought of as providing a *comparison* between a small closed economy and an (otherwise identical) small open economy. It is not a result about the effects on a given economy of the process of opening up to trade, full account being taken of what happens during the transition from the autarky state to the free trading state. But the same is true, it must be noted, of the textbook demonstrations of the gain from trade, in a 'Ricardian' framework, with which the reader is familiar.

While 'factor price equalization' is most often discussed within the Heckscher–Ohlin–Samuelson (HOS) framework, it is of interest to consider whether free trade in all commodities will bring about real wage rate and interest rate equalization in the type of model considered here. If all the freely trading economies have the same available choice of techniques, in a constant-returns-to-scale and homogeneous labour world, then it is certainly true that, if they all have the same rate of interest, they will all have the same set of relative prices. But the converse does *not* hold, when there is a choice of techniques; all the economies could face the same set of relative commodity prices and yet have different interest rates and real wage rates. Hence free trade in all commodities does not entail wage and interest equalization, even when all the economies have the same technical possibilities and are incompletely specialized. (The same negative conclusion holds, even when there is no choice of technique, if there are non-traded commodities.)

(On the pattern of trade and the gain from trade see Mainwaring 1974; Samuelson 1975; Steedman and Metcalfe 1973a, 1979; Steedman 1979a. On interest rate (non-) equalization see Mainwaring 1976, 1978; Samuelson 1975; Steedman and Metcalfe 1973b.)

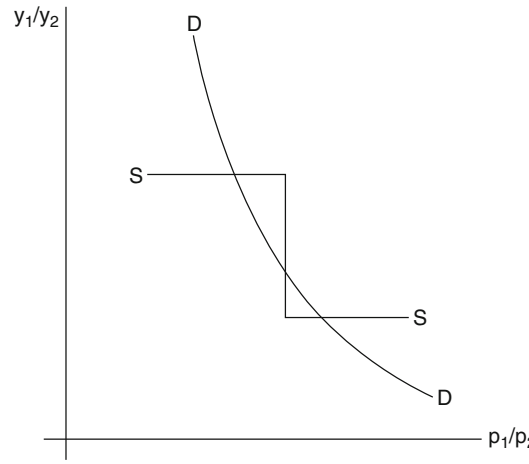
Land, Labour and a Positive Interest Rate

We now turn to the much-loved HOS model of international trade, in which two countries produce the same two commodities, using the same two primary inputs (which are in fixed supply) and having the same, constant-returns-to-scale technology. The primary inputs are qualitatively the same in both countries, fully mobile within each economy but completely immobile between them. There are no factor-intensity reversals, there is completely free trade and all consumers, in both countries, share a common homothetic preference map (so that consumption proportions depend only on the commodity price ratio, being quite independent of income distribution). If the two primary inputs are homogeneous land and

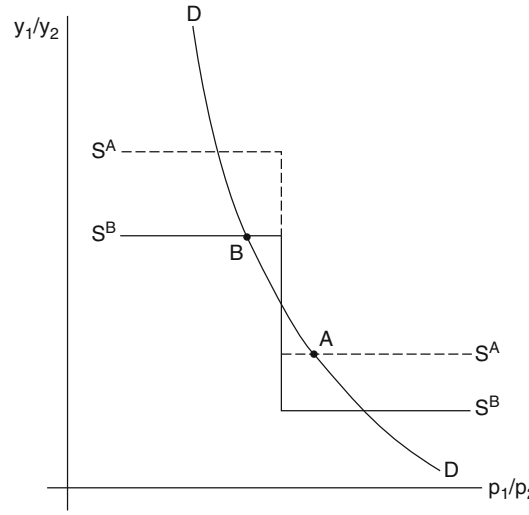
homogeneous labour, and if there are no produced inputs (capital goods) of any kind, then the HOS theorem on the pattern of trade (in both its price and quantity forms), the factor price equalization theorem, the Stolper-Samuelson theorem and the Rybczynski theorem are all logically valid theorems.

Suppose now that, retaining all the other assumptions, we allow the two consumption commodities also to be capital goods, being necessary inputs to the various productive processes. What difference does this introduction of produced inputs make to the standard theorems? None whatever! It is now more appropriate to think of land/labour intensities in production in terms of *total* (direct and indirect) uses of land and labour but, since the intensity ranking of commodities in these total terms is necessarily the same as that in direct terms, this introduces no really significant difference from the model without produced inputs. Thus far then, produced inputs make no difference. But the position changes as soon as we allow not only for the presence of such produced inputs but also for a given, *positive* rate of interest on the value of those inputs (circulating capital goods). The presence of a positive interest rate does not alter the fact that the relative price of the land-intensive commodity will be a monotonically increasing function of the rent/wage ratio. But, as was pointed out above, it does mean that an increase in the rent/wage ratio is not necessarily associated with a fall in the land/labour ratio; it then follows that, if land and labour are always fully employed, an increase in the relative price of the land-intensive commodity may be associated with a *fall* in its net output. In Fig. 2, which relates to a single economy, y_i is the net product of i,p_i is the price of i, SS is the full employment 'relative supply curve' and DD is the 'relative demand curve' derived from the common homothetic preference map; the figure illustrates the case of a 'perverse' supply response. It will be seen at once that such a supply response immediately gives rise to the possibility of multiple equilibria, the 'first' and 'third' equilibria both being stable.

Let two economies, A and B, have the same positive rate of interest; let A be relatively better endowed with land and let commodity 1 be the land-intensive commodity. Figure 3 extends Fig. 2 to this case, S^iS^i being the full employment relative supply curve for economy i. Suppose that point A represents A's autarky equilibrium, while point B represents B's. At every (p_1/p_2) lying *between* the autarky price ratios, S^AS^A and S^BS^B both lie on the same side of DD; hence no such price ratio can be an equilibrium terms of trade. The terms of trade lie *outside* the autarky price range. Whether the international equilibrium is found to the left of B or to the right of A, economy A (which is well endowed with land) will be



Foreign Trade, Fig. 2



Foreign Trade, Fig. 3

exporting commodity 1 (which is the land-intensive commodity). Thus the HOS quantity theorem holds good. Yet the HOS price theorem, which is sometimes thought rather trivial, as actually *false* here. Since A has the higher autarky (p_1pp_2), it has the higher autarky rent/wage ratio, so that A is exporting the commodity which uses intensively A's relatively *expensive* factor under autarky. Notice also that if international equilibrium is found to the left of B, (p_1/p_2) will have fallen, with trade, in economy A and thus the wage/rent ratio will have *risen*; in fact trade will have *benefited* A's relatively scarce factor (labour), contrary to the usual HOS prediction.

If A and B have the *same* positive interest rate, as above, they have the same relationship between (p_1/p_2) and the rent/wage ratio; it is thus not surprising that free trade will equalize rents and wages (with incomplete specialization) and that the Stolper–Samuelson theorem also holds good. If A and B have *different* positive interest rates, however, almost everything collapses. The exception is the Rybczynski theorem and it is important to understand why. All 'capital theoretic' problems for HOS theory reduce in the end to the fact that relative commodity prices vary with the rate of interest – but relative prices are fixed *by assumption* in the Rybczynski theorem, so that that theorem must be immune to such problems.

Consider now a single, small economy of the kind discussed immediately above. In the presence of a positive interest rate, the price ratio at which a switch of techniques takes place will not be equal, in general, to the physical rate of transformation between the two net outputs. It follows that, when we compare the small open economy with an otherwise identical autarkic economy, we find that the value of consumption in the small open economy, at the given international prices, may be either greater than or less than the corresponding value in the autarkic economy. The 'comparative static' gain from trade may be either positive or negative.

In the land and labour model, then, the presence of produced inputs makes no difference per se. But a positive rate of interest on their value does make a difference to some (but not all) HOS theorems, if it is the same in both countries, while a difference in interest rates undermines all the standard HOS results, other than the Rybczynski theorem. For the single, small, open economy the presence of a positive interest rate means that the 'comparative' gain from trade can be positive or negative.

(For the closed economy background see Metcalfe and Steedman 1972; Montet 1979; for the trade theory applications Samuelson 1975; Steedman and Metcalfe 1977; for the gain from trade Metcalfe and Steedman 1974; Samuelson 1975.)

F

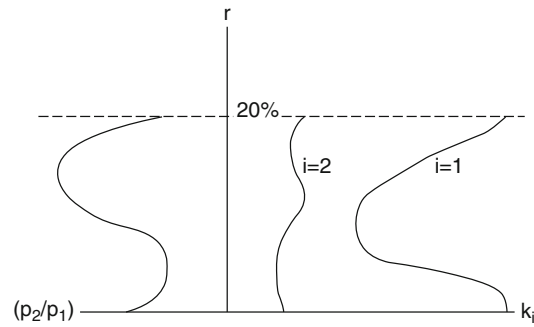
Labour and Capital

In the typical textbook presentation of HOS theory the two 'factors' in given supply are not labour and land, as above, but labour and 'capital'. (Although Samuelson (1948, 1949) was careful to stipulate labour and land.) Yet that typical presentation suggests no immediate connection between the two produced commodities and the physical composition of the capital stock, despite the fact that 'capital goods' are, by definition, produced means of production! Indeed, one interpretation of most textbook theory is that 'capital' is simply a misnomer for land, the problems of capital theory being evaded by a simple misuse of terms. Alternatively (and more favourably), the 'given capital supply' can be interpreted to mean that the total *value* of capital goods must always be equal to – or, at least, not greater than – an exogenously given value. An immediate difficulty with this interpretation is that, since relative autarky prices differ between the two economies, the very *ranking* of the two countries' capital/labour endowments ratios may depend on which standard of value is used to measure capital. And what does it *mean* economically to suppose that total capital value is given in terms of one standard and yet, necessarily, is *not* given in terms of all other possible standards (since relative commodity prices are to be determined endogenously)? Even if we ignore these questions – which there is no justification for doing – we know from capital theory that value capital/labour ratios need not be related inversely or, indeed, even

monotonically to the rate of interest. This, of course, immediately suggests that some of the HOS theorems may be at risk. Moreover, it can be shown that in a model with many produced inputs, the price ratio between any two particular commodities need not be monotonically related to the rate of interest, *even when* one of the two commodities is always more value capital-intensive than the other. But if neither the capital/labour ratios nor the relative commodity prices need be monotonically related to the rate of interest – even in the absence of factor-intensity reversals – then it will at once be clear that HOS theorems (other than the Rybczynski theorem) cannot be logically valid when one of the two factors is a 'given value of capital'. This stems fundamentally from the simple fact that Wicksell clearly stated many years ago:

> Whereas labour and land are measured each in terms of its own *technical* unit ... capital ... is reckoned, in common parlance, as a sum of *exchange value* – whether in money or as an average of products. In other words, each particular capital-good is measured by a unit extraneous to itself. [This] is a theoretical anomaly which disturbs the correspondence which would otherwise exist between all the factors of production. ([1901] 1967, p. 149)

To illustrate the above negative conclusions, we may use an example in which there are two consumption commodities (two kinds of 'corn'), each producible by means of many alternative types of machine. The consumption commodities are tradeable but the machines are not. Full numerical details of this example can be found in Metcalfe and Steedman (1973); here we confine ourselves to the diagrammatic presentation of Fig. 4, in which k_i is the value capital/labour ratio involved, directly and indirectly, in the production of the ith consumption commodity, expressed in terms of the first consumption commodity. It will be seen on the right of Fig. 4 that neither k_1 nor k_2 is monotonically related to r but that $k_1 > k_2$ at all r; on the left we see that, the absence of factor-intensity reversal notwithstanding, the price ratio (p_1/p_2) is not monotonically related to r. It follows at once that the 'factor price' equalization theorem, the Stolper-Samuelson theorem and the price form of the HOS theorem on the pattern of trade



Foreign Trade, Fig. 4

are *not* of general logical validity. But if the pattern of trade theorem is not valid in its price form then it will not be valid in its quantity form either, even if it is the case (which it may not be) that the economy with the higher capital/labour endowment ratio has the lower autarky interest rate.

When produced inputs are introduced into HOS theory in the form that one of the two 'factors' is taken to be a given total value of capital, that theory simply disintegrates. This is so notwithstanding the apparent denial of this negative conclusion by Ethier (1979), who states that 'The central message ... is simple. The four basic theorems of the modern theory of international trade ... are insensitive to the nature of capital' (p. 236). In fact Ethier's paper constitutes a striking confirmation of our negative conclusion, because in order to maintain the *appearance* that capital has no influence on HOS trade theorems, Ethier finds himself compelled to *replace* the familiar theorems, which predict trade outcomes on the basis of exogenous data, by entirely different theorems, which merely describe trade outcomes in terms of trade equilibrium prices, etc.

(For the example used in this section, see Metcalfe and Steedman 1973; on Ethier's conjuring with HOS theorems, see Metcalfe and Steedman 1981.)

Growth, International Investment and Transitions

To focus on the role of capital goods in trade and in trade theory is, implicitly, to direct attention

also to such matters as growth (capital accumulation), international investment and transitions between steady growth paths. Since the typical trading economy uses many produced inputs – some traded and some not – accumulates capital goods and experiences (often embodied) technical change, along both quantitative and qualitative dimensions, the ideal theory of trade would be able to handle all these closely related issues, in a manner which was both informative and simple. Needless to say, such an ideal theory is not available; international trade theory in these respects can, in the long run, be no more advanced than the general theory of accumulation and technical progress. The preceding discussion can, however, serve to warn us that growth models in which there is a single, physical capital good can almost certainly not be readily generalized to the many capital good case and are thus of *very* limited interest. It is also useful to note, as a simple matter of fact, that while the number of countries in the world is of the order of 200, the number of distinct commodities – when defined at the level of detail relevant to careful value theory – runs into millions. This both tells us that incomplete specialization must be the rule and directs our attention to economic growth models in which the number of commodities can be arbitrarily large; the von Neumann model perhaps deserves to be used more extensively by trade theorists than it has been, its very abstract nature notwithstanding.

When thinking of capital accumulation, the international economist will naturally pay considerable attention to the role of international investment. Here it is most important to recognize that, although they are often connected in practice, there is a perfectly clear – indeed a sharp – distinction between international investment as a flow of finance, on the one hand, and trade in physical capital goods, on the other. This is obvious enough perhaps when stated explicitly but it is to be noted that the idea of a 'factor' capital, conceived of as a sum of value, in fact makes it dangerously easy to confuse financial flows with capital goods flows. The trade theorist would do well to avoid the concept of a 'quantity of capital' altogether, referring only to stocks and flows of specified capital goods, on the one hand, and to international flows of finance, on the other. Such a practice would not only make it easier to avoid capital theory traps but would also facilitate thought about the badly needed integration of pure trade theory with international monetary economics.

We turn now to the question of 'transitions'. Consider first a closed economy whose homogeneous land and homogeneous labour are allocated between strawberry production and raspberry production. No produced inputs are used – not even strawberry and raspberry plants! (Which reminds us, incidentally, of just *how* strained is any picture of direct production of consumption commodities by primary inputs.) If free trade should suddenly become possible, at terms of trade different from the autarky price ratio, there is no difficulty at all in reallocating the land and labour to the newly desired output pattern. The 'transition' from the autarky steady-state to the with-trade steady-state is problem free and can be achieved instantaneously. By contrast, consider now the analogous 'transition' for an economy which does use produced inputs. Except by a complete fluke, the economy's industries will use the various produced inputs in different proportions from one another and it will now *not* be possible to change to the free trade pattern of output instantaneously. Since the production of the produced inputs takes time, there will have to be a 'transitional' period, during which the physical composition of the economy's aggregate capital stock is adjusted to the new output pattern. Just how long this period will be depends, of course, on how different the input requirements are as between industries, on whether or not some previously used capital goods simply have to be scrapped, on how many of the capital goods are tradeable and how many non-tradeable, etc. Changing the pattern of net output is a far more complicated process in an economy using produced inputs. This issue is avoided in textbook discussions of the gain from trade *and* in the 'comparative dynamics' results given above. Yet it can hardly be denied that the issue is important in many trade policy applications and in many day-to-day debates about trade protection, industries which are under increased international competitive pressure, and so on. It is therefore important that trade theorists should

develop *explicit* analyses of transitional processes in the presence of produced inputs. At the same time, however, it would be quite wrong simply to dismiss 'comparative dynamic' results showing that a 'loss from trade' is possible, merely on the grounds that (by definition) they do not take account of transitions. The traditional comparisons of a world of autarky economies with a world of trading economies are designed to show that the with-trade state of the world (which we observe) is preferable to the autarky state (which is purely hypothetical). For the purpose of such an abstract, *hypothetical comparison*, the analysis of transitions would have no significance and it is indeed the purely 'comparative' analysis which is relevant. (There is, of course, no inconsistency in saying also that a transitional analysis is relevant for the study of an actual economy considering the possibility of, say, changing its tariff structure.)

(A trade theory application of the von Neumann model is given in Steedman 1979c, for a single, small economy; growth in a two country world is discussed by Parrinello 1979. On transitions see Metcalfe and Steedman 1974; Smith 1979.)

Conclusion

Sufficient reason has perhaps been given above to justify the rather general conclusion that when one finds trade theorists referring to 'capital' one should immediately be 'on guard'. The presence of produced inputs, with a positive rate of interest on their value, does make a considerable difference to the logical coherence of HOS theory, as has been seen in some detail above. Moreover, 'textbook' Ricardian trade theory, which appears to make no reference to 'capital' at all, ought to make such reference and, if it did, would discover that here again the presence of a positive interest rate makes it far harder to reach any clear cut, logically valid theorems. In seeking to develop a trade theory which *does* give central importance to capital goods and hence to profits, accumulation and technical progress (e.g. Steedman 1979a) one must expect that simple results may not be abundant. And one must recognize that the assumptions which make growth theory relatively easy, such as constant returns to scale and the absence of land, themselves do violence to the complex realities of international trade. (Its many shortcomings notwithstanding, HOS theory is right to stress the importance of land and labour endowments, even while it is wrong to take them to be qualitatively homogeneous and fully employed.) The role of capital goods is by no means the only important issue in trade theory and recognition of that role certainly makes trade theory more difficult. But can these be good reasons for ignoring capital goods, when that theory is intended to aid our understanding of a world in which produced inputs are, in fact, centrally important?

See Also

- ▶ Heckscher–Ohlin Trade Theory
- ▶ International Trade

Bibliography

Ethier, W.J. 1979. The theorems of international trade in time-phased economies. *Journal of International Economics* 9(2): 225–238.

Mainwaring, L. 1974. A neo-Ricardian analysis of international trade. *Kyklos* 27(3): 537–553. Reprinted in Steedman (1979b).

Mainwaring, L. 1976. Relative prices and 'factor price' equalisation in a heterogeneous capital goods model. *Australian Economic Papers* 15(26): 109–118. Reprinted in Steedman (1979b).

Mainwaring, L. 1978. The interest rate equalisation theorem with nontraded goods. *Journal of International Economics* 8(1): 11–19. Reprinted in Steedman (1979b).

Metcalfe, J.S., and I. Steedman. 1972. Reswitching and primary input use. *Economic Journal* 82: 140–157. Reprinted, with minor corrections, in Steedman (1979b).

Metcalfe, J.S., and I. Steedman. 1973. Heterogeneous capital and the Heckscher-Ohlin-Samuelson theory of trade. In *Essays in modern economics*, ed. J.M. Parkin. London: Longman. Reprinted in Steedman (1979b).

Metcalfe, J.S., and I. Steedman. 1974. A note on the gain from trade. *Economic Record* 50: 581–595.

Metcalfe, J.S., and I. Steedman. 1981. On the transformation of theorems. *Journal of International Economics* 11(2): 267–271.

Montet, C. 1979. Reswitching and primary input use: A comment. *Economic Journal* 89: 642–647.
Parrinello, S. 1979. Distribution, growth and international trade. In Steedman (1979b).
Pasinetti, L.L. 1977. *Lectures on the theory of production*. London: Macmillan.
Quarterly Journal of Economics. 1966. Symposium on paradoxes in capital theory. *Quarterly Journal of Economics* 80(4): 503–583.
Ricardo, D. 1817. In *Principles of political economy and taxation*, ed. P. Sraffa. Cambridge: Cambridge University Press, 1951.
Samuelson, P.A. 1948. International trade and the equalization of factor prices. *Economic Journal* 58: 163–184.
Samuelson, P.A. 1949. International factor-price equalization once again. *Economic Journal* 59: 181–197.
Samuelson, P.A. 1975. Trade pattern reversals in time-phased Ricardian systems and intertemporal efficiency. *Journal of International Economics* 5(4): 309–363.
Smith, M.A.M. 1979. Intertemporal gains from trade. *Journal of International Economics* 9(2): 239–248.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Steedman, I. 1979a. *Trade amongst growing economies*. Cambridge: Cambridge University Press.
Steedman, I. (ed.). 1979b. *Fundamental issues in trade theory*. London: Macmillan.
Steedman, I. 1979c. The von Neumann analysis and the small open economy. In Steedman (1979b).
Steedman, I., and J.S. Metcalfe. 1973a. On foreign trade. *Economia Internazionale* 26(3–4): 516–528. Reprinted in Steedman (1979b).
Steedman, I., and J.S. Metcalfe. 1973b. The non-substitution theorem and international trade theory. *Australian Economic Papers* 12(21): 267–269. Reprinted in Steedman (1979b).
Steedman, I., and J.S. Metcalfe. 1977. Reswitching, primary inputs and the Heckscher-Ohlin-Samuelson theory of trade. *Journal of International Economics* 7(2): 201–208. Reprinted in Steedman (1979b).
Steedman, I., and Metcalfe, J.S. 1979. The golden rule and the gain from trade. In Steedman (1979b).
Wicksell, K. 1901. *Lectures on political economy*, vol. 1. Trans. E. Classen, ed. Lionel Robbins. London: G. Routledge and Sons, 1934.

Foreign Trade Multiplier

K. J. Coutts

Though its ancestry may be traced to certain ideas of the Mercantilist School, the foreign trade multiplier appears in modern form in a textbook (Harrod 1933) written three years before Keynes's *General Theory*. There in its simplest form Harrod introduced the familiar equation for the determination of the national income flow consistent with balance in the current account of the balance of payments – that the national income is equal to the product of the volume of national exports and the reciprocal of the average propensity to import. In this form it emphasized the equilibrating role of income over price variations in determining balance of payments adjustment, and the importance that trade performance played in determining the level of activity at which external balance would be achieved.

By the 1939 edition Harrod was expounding his foreign trade multiplier as an application of Keynes's theory of effective demand by incorporating domestic investment and the savings propensity within the analysis.

After World War II great strides were achieved in integrating relative prices and income into the analysis of balance of payments adjustment (Laursen and Metzler 1950; Harberger 1950; Meade 1951). The foreign trade multipler was subsumed in a synthesis of balance of payments theory which became known as the absorption approach (Alexander 1952). Theory became organised around the identity for the current account of the balance of payments expressed as the difference between the national income and domestic expenditure or absorption per period.

A further generalization was to analyse the foreign trade multiplier by including the foreign repercussions of changes in spending and income which feed back onto the home country, initiating the change in autonomous expenditure. An essentially identical treatment of the multiplier could be applied either to a multi-sectoral, regional or country analysis of the determination of income by autonomous spending in any single unit of the system. Thus the focus changed from the relationship between the national economy and the rest of the world to the inter-relationships of a multi-country world economy in which trade shares and absorption propensities played an important role in determining the global and distributional generation of income (Goodwin 1949; Metzler

1950). Within this framework the post-war problems of international trade and growth could be addressed, where for the world as a whole, the closed economy assumption of theory is literally (and almost uniquely) true.

By the middle of the 1950s the major theoretical developments of multiplier analysis in open-economy macroeconomics had been established.

In contrast with the pure theory of international trade, set in a context ensuring full employment of resources and concerned mainly with factors which determine efficient patterns of trade, the foreign trade multiplier is directly relevant to the determination of the level of output and employment between nations. The extension of effective demand theory into international economics has given greater prominence to income and employment changes rather than relative price changes at full employment in balance of payments adjustment. Whether under fixed or floating exchange rate regimes the foreign trade multiplier indicates how changes in *net* exports, i.e. the excess of export growth over the growth of import penetration, affect both the balance of payments on current account and the degree of domestic capacity utilization.

For all countries which trade significantly with one another, the foreign trade multiplier serves to remind one that the pursuit of full employment and economic growth policies by governments must be accompanied by policies which ensure that they are matched by satisfactory balance of payments performance. This implies that the growth of net exports makes a significant contribution to the growth of national income.

A problem of paramount importance for any system of international settlements adopted between nations is how national fiscal and monetary policies, enacted largely independently of one another, can achieve mutually reconcilable balance of payments outcomes without forcing at least some nations to abandon desirable domestic objectives. The foreign trade multiplier, appropriately more complex than Harrod's original formulation, provides an illuminating framework for addressing this major question.

See Also

- ▶ Absorption Approach to the Balance of Payments
- ▶ International Trade
- ▶ Transfer Problem

Bibliography

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

Goodwin, R.M. 1949. The multiplier as matrix. *Economic Journal* 59: 537–555.

Harberger, A.C. 1950. Currency depreciation, income, and the balance of trade. *Journal of Political Economy* 58(1): 47–60.

Harrod, R.F. 1933. *International economics*, Cambridge economic handbooks. Cambridge: Cambridge University Press; revised ed, 1939.

Laursen, S., and L.A. Metzler. 1950. Flexible exchange rates and the theory of employment. *Review of Economics and Statistics* 32(4): 281–299.

Meade, J.E. 1951. *The theory of international economic policy. Vol. I: The balance of payments*. London: Oxford University Press.

Metzler, L.A. 1950. A multiple region theory of income and trade. *Econometrica* 18(4): 329–354.

Forests

P. A. Neher

Traditional forestry economics has been chiefly concerned with wild or cultured forests as commercial, agricultural, enterprises. For these, net economic benefits stem from the harvested timber and the objective is to calculate the optimal pattern of harvesting over time. While there is a venerable literature on the standing, *in situ*, values of trees (see J. Nisbet's entry in *Palgrave* (1912), Vol. II, pp. 113–18), these have been incorporated only recently in formal optimizing models.

The early work of Martin Faustmann (1849) is noteworthy for its originality and for providing the correct solution to the central problem of the optimal rotation period for a sequence of harvests. For modern interpretations see Gaffney (1960), Pearse (1967) and Gregory (1972).

Stripped to its essentials, the Faustmann problem is to maximize the present value of bare land (V) which will support an indefinite sequence of harvests ($n = 1,2,\ldots,\infty$). The trees grow in value (net of harvesting costs) as they mature according to $P(T)$. $P(T)$ is nil for a while, then increases at a decreasing rate, reaching a maximum then falling (as rot sets in). Then harvest is instantaneous at T. Regeneration costs are nil. The problem is to

$$\max_{\{T\}} V(T) = \sum_{n=1}^{n=\infty} D(T)^n P(T). \tag{1}$$

The discount factor, applied to the harvest when the trees are T years old, is $D(T) = \exp(-T)$ for continuous compounding and $D(T) = (1 + i)^{-t}$ for annual compounding. For example, the contribution to $V(50)$ from the third harvest is by fifty-year-old trees and is computed as $D(50)^3 P(50)$.

The problem is solved by first recognizing that $D(T)^n P(T)$ in (1) is a declining, infinite, geometric series having a finite sum. Thus

$$V(T) = [D(T)/(1 - D)]P(T) \tag{2}$$

$V(T)$ at first rises, then falls as T is extended beyond the interval while $P(T) = 0$. If T is only a little greater than this, V is still nearly zero since $P(T)$ is small. If T is long, $P(T)$ is large but the expression in square brackets is small. $V(T)$ is a maximum when $V'(T^*) = 0$ or

$$P'(T^*)/P(T^*) = -D'(T^*)/[D(T^*)(1 - D(T^*))] \tag{3}$$

This solves (1) for T^*, the age of *financial maturity* (Duerr et al. 1956). It is the (by now) famous *Faustmann* (1849) *Formula*.

For continuous compounding, $D'(T)\ /D(T))$ equals $(-r)$ and

$$P'(T^*)/P(T^*) = r(1 - \exp rT^*)^{-1} \tag{4}$$

Substituting this result back into (2) gives

$$P'(T^*) = r(P(T^*) + V(T^*)) \tag{5}$$

The RHS of (5) contains the maximized present value of the bare land, $V(T^*)$. This is called the *land-expectation* value. Added to this is the optimal *stumpage* value of the trees, $P(T^*)$. The two values together are the value of the land plus the value of the trees, both evaluated at T^*. This sum multiplied by r is the momentary interest payment earned by selling the cut trees *plus* the bare land, then putting the money 'in the bank' earning interest at the rate r. The LHS of (5) is the momentary increase in the value of the trees if left on the stump to grow in the ground. The time to cut, T^*, is when the marginal increment in accrued wealth is the same with the trees 'in the ground' as 'in the bank'.

A special case of the Faustmann result is obtained by supposing that the land has no value other than supporting the first growth of trees. Then, from (5), the *present net worth* of the trees alone is maximized when

$$P'(T^*) = rP(T^*) \tag{6}$$

This result is associated with 1. Fisher (1930), Hotelling (1925), von Thünen (1826). It has been called the *Fisher Rule*. When the trees are younger than T^*, they grow in value in the ground faster than their value if cut would increase in the bank. And vice versa after T^*. The moment to cut and sell is the moment of indifference. Since $P'(T^*)$ is greater for smaller T, the Fisher rule signals a later cut than does the Faustmann formula. This is because the Fisher Rule does not capture the impatience of a forester to cut out the old growth to make room for younger and faster growing trees.

Another special case of the Faustmann formula is obtained by supposing that the relevant rate of interest (r) is zero. Then, using I'Hôpital's rule in (4)

$$\lim_{r \to \infty} P'(T^*)/P(T^*) = 1/T^*. \tag{7}$$

Since younger trees grow faster in value, (7) signals a later cut than does (3) with r positive.

The rule in (7) can be obtained directly by maximizing the *mean annual increment* (MAI) with respect to the age of trees when harvested.

F

$$\max_{\{T\}} MAI = P(T)/T.$$

This objective has been widely accepted by professional foresters since the late 18th century (Osmaston 1968). The harvesting rule, expressed in (7), is elegant and parsimonious of information requirements. In practice, $P(T)$ is usually approximated by the volume of merchantable timber contained in a tree. Hence, the rule has no economic content except in so far as it may serve as a practical rule-of-thumb approximation of a necessary condition for an economic optimum.

Samuelson (1976) provides a concise and readable comparison of these versions of the Faustmann formula. He concludes that Faustmann's original formulation is the 'correct' one for maximizing the social contribution of forests. It should, however, be interpreted as 'pure theory', subject to the austere conditions imposed to reveal the kernel of the problem.

The Faustmann formula has been enriched to encompass various kinds of silviculture including artificial regeneration, thinning, fertilization, insect and fungus control. Felling, yarding, bucking and transportation costs have also been explicitly included in net revenue functions. Clark (1976), Ledyard and Moses (1976) and Heaps (1981) provide examples. These are exercises in operations research and serve as guides to formulating cost-benefit studies on a case-by-case basis. A general consideration is that costs incurred early in a rotation (artificial regeneration, for example) will be compensated by earlier cuts or enhanced net current values many years later. Calculations made for coastal forest land in northwest United States (southwest Canada) compute the economic value of re-established forests after clear-cutting by approved methods (Smith 1978). The reference forest is a naturally regenerated wild stand of mixed low and high valued trees harvested after 76 years. The enhanced forest is planted with genetically improved high valued species and the harvest is accelerated after a (non-commercial) thining at 15 years. The additional cost of planting and thinning is $580 per hectare, increasing the *current* value of the harvest by $2794. But the increase in *present* value calculated at a 5 per cent p.a. interest rate is only $78 per hectare. Plantation forests on good, flat land in warmer climates have shorter economic rotations so thinnings and other kinds of silviculture may have net commercial value.

Traditional forest management contemplates *sustained yield* from a *regular*, or *normal, forest* which encompasses sufficient geographic scope to sustain a continuous harvest of trees of the desired age. This requires x even-age, equal-area, stands of trees ranging in age from zero to $(x–1)$. A stand is cut when the trees are x years old. Each stand is to be harvested during a year of an x year *rotation*. If the interest rate is zero in the Faustmann formula, the x is T for the maximum MAI to obtain *Maximum Sustained* (in perpetuity) *Yield* (MSY) in terms of timber volume. Rotations are shorter for positive interest rates and fewer even-age stands are required in a management unit.

The regular forest is the forester's ideal. But the initial condition of a forest may be characterized by irregular age distributions. *Ad hoc* formulae for conversion to a regular forest have been proposed by Hanzlik (1922) and others. See Hennes et al. (1971). Naudial and Pearse (1967) model conversion as an economic problem. Using a linear programme, they solve for an optimal rotation period (T^*), *given* a time period for conversion. But V is seen to rise as the period is extended, implying that the regular forest is not optimal. Heaps (1981) and Heaps and Neher (1979) provide the most general treatment of the problem using the Maximum Principle for processes with an endogenous delay (between regeneration and cutting). Variable ('u'-shaped average) costs of harvest are allowed for. The Forestry Maximum Principle leads to a dynamical system of functional differential equations. It is seen that a convergent harvesting policy yields the Faustmann T^* for a regular forest. Global (asymptotic) stability has not been proved but it seems likely that conversion to a regular forest over a period of several rotations is practically optimal.

Other recent contributions allow for stock uncertainty and for standing values of existing tree stocks. Reed (1984) shows that the expected valuation of sustained yield is maximized by

planning to cut before the Faustmann T^*. The early cut forestalls the possibility of catastrophic loss. Also see Kao (1984), Martell (1980), and Reed and Errico (1985).

The standing value of trees has long been recognized in the literature (Palgrave 1912). Indeed, some notable modern forests are extant relics of ancient game preserves (Epping Forest, London, for example). Vast forested areas of the United States are administered by the Forest Service under multiple-use mandates which include preservation of wildlife habitat and recreational values. Recent widespread clearings of tropical rainforest have focused attention on the value of forested areas to sustain (sometimes unique) fauna and flora. Hartman (1976) and Neher (1976) provide early examples of theoretical frameworks for including standing values in intertemporal optimizing models.

The inclusion of standing values presents both interesting theoretical problems and challenging management difficulties. The kernel of the theoretical problem is compactly exposed by absorbing the age-class demographic structure of the forest into a simple, biomass aggregate of trees (B). Let B grow naturally according to $G(B) > 0$ with $G(B)$ a maximum at $B_M > 0$, $G(0) = G(\overline{B}) = 0$ and $G''(B)$. $\overline{B}$ represents the climax forest. Cardinal social benefits depend upon the harvest flow (H) and upon the standing stock (B). In short, $U = H(H, B)$. Let $U_H, U_B > 0$ and $U_{HH}, U_{BB} < 0$. However $U_{BH} = U_{HB}$ is not signed *a priori*. The value of (V) of the harvesting plan is the (undiscounted) sum (integral) of these U's over the planning interval $[0, T]$. V is to be maximized subject to constraints that $B(0) = B_0$ (the original state of the forest is given) and that $''B = G(B) - H$ (nature's own 'budget constraint') is satisfied. Thus, $U = U(G(B) - B, B)$. It is well known that the necessary conditions for an optimal programme are not sufficient if U () is not jointly concave in (B, B). In this case, sufficiency is not guaranteed unless marginal enjoyments of H and B are independent of each other ($U_{HB} = U_{BH} = 0$). This phenomenon was originally identified by Kurz (1968) in neoclassical growth where multiple equilibria were identified. Cropper (1976) provides a more recent example in the optimal control of pollution. Additional sources of multiple equilibrium are introduced by considering many (natural and produced) capital stocks (Heal 1982) and by discounting future benefits (Cass and Shell 1976). Much work has yet to be done before there is general theoretical understanding of optimal multiple-use forestry.

It is not surprising that practical, implementable, multiple-use models have not been devised. Single-use (timber value) commercial models are conceptually more straightforward, incorporating (if at all) standing values as exogenous constraints on the available commercial area. These are typically linear programming models. See Johnson and Scheurman (1977) for an evaluative survey. The discrete maximum principle (Halkin 1966) offers a promising alternative (Lyon and Sedjo 1983). Large-scale programming models have been implemented with arguable success (Timber Resources Allocation Model (Timber RAM); Navon 1971; FORPLAN, Johnson et al. 1980).

See Also

- ▶ Faustmann, Martin (1822–1876)
- ▶ Natural Resources

Bibliography

Cass, D., and K. Shell. 1976. *The Hamilton approach to dynamic economics*. New York: Academic.

Clark, C.W. 1976. *Mathematical bioeconomics*. New York: Wiley.

Cropper, M.L. 1976. Regulating activities with catastrophic environmental effects. *Journal of Environmental Economics and Management* 3: 1–15.

Duerr, W.A., Fedkiw, J., and Guttenburg, S. 1956. *Financial maturity*. US Department of Agriculture. Technical Bulletin No. 1146.

Faustmann, M. 1849. *On the determination of the value which forest land and immature stands pose for forestry*. English translation in *Martin Faustmann and the evolution of discounted cash flow*. ed. M. Gane. Oxford Institute Paper 42, 1968.

Fisher, I. 1930. *The theory of interest*. New York: Macmillan.

Gaffney, M.M. 1960. *Concepts of financial maturity of timber and other assets*, Agricultural economics information series, vol. 62. Raleigh: North Carolina State College.

F

Gregory, G.R. 1972. *Forest resource economics*. New York: Ronald Press.
Halkin, H. 1966. A maximum principle of the Pontryagin type for systems described by non-linear difference equations. *SIAM Journal on Control* 4: 90–111.
Hanzlik, E.J. 1922. Determination of the annual cut on a sustained basis for virgin American forests. *Journal of Forestry* 20: 611–625.
Hartman, R. 1976. The harvesting decision when a standing forest has value. *Economic Inquiry* 16: 52–58.
Heal, G. 1982. The use of common property resources. In *Explorations in natural resource economics*, ed. V.K. Smith and J.V. Krutilla. Baltimore: Johns Hopkins Press.
Heaps, T. 1981. The qualitative theory of optimal rotations. *Canadian Journal of Economics* 14: 686–699.
Heaps, T., and P.A. Neher. 1979. The economics of forestry when the rate of harvest is constrained. *Journal of Environmental Economics and Management* 6: 279–319.
Hennes, L.C., Irving, M.J., and Navon, D.I. 1971. *Forest control and regulation*. USDA Forest Service Research Note PSW–231. Berkeley.
Hotelling, H. 1925. A general mathematical theory of depreciation. *Journal of the American Statistical Association* 20: 340–353.
Johnson, K.N., D.B. Jones, and B. Kent. 1980. *A user's guide to the forest planning model (FORPLAN)*. Ft. Collins: Land Management Planning, USDA Forest Service.
Johnson, K.N., and H.L. Scheurman. 1977. *Techniques for prescribing optimal timber harvest and investment under different objectives*, Forest science monographs, vol. 18. Washington, DC: Society of American Foresters.
Kao, C. 1984. Optimal stocking levels and rotation under uncertainty. *Forest Science* 30: 921–927.
Kurz, M. 1968. Optimal economic growth and wealth effects. *International Economic Review* 9: 348–357.
Ledyard, J., and L.N. Moses. 1976. Dynamics and land use: The case of forestry. In *Public and urban economics*, ed. R.E. Frieson. Lexington: D.C. Health.
Lyon, K.S., and R.A. Sedjo. 1983. An optimal control theory model to estimate the regional long-term supply of timber. *Forest Science* 29: 798–812.
Martell, D. 1980. The optimal rotation of a flammable forest stand. *Canadian Journal of Forest Research* 10: 30–34.
Navon, D.J. 1971. *Timber RAM*, USDA Forest Service Research Paper PNW–70. Berkeley: Pacific Southwest Forest and Range Experimental Station.
Naudial, J.C., and P.H. Pearse. 1967. Optimizing the conversion to a sustained yield. *Forest Science* 13: 131–139.
Neher, P.A. 1976. Democratic exploitation of a replenishable resource. *Journal of Public Economics* 5: 361–371.
Osmaston, F.C. 1968. *The management of forests*. London: Allen & Unwin.
Palgrave, R.H.I. (ed.). 1912. *Dictionary of political economy*, vol. II, 2nd ed. London: Macmillan.
Pearse, P.H. 1967. The optimum forest rotation. *Forestry Chronicle* 43: 178–195.
Reed, W.J. 1984. The effect of risk of fire on the optimal rotation of a forest. *Journal of Environmental Economics and Management* 11: 180–190.
Reed, W.J., and D. Errico. 1985. Optimal harvest scheduling at the forest level in the presence of the risk of fire. *Canadian Journal of Forest Research* 15: 680–687.
Samuelson, P. 1976. Economics of forestry in an evolving society. *Economic Inquiry* 14: 466–492.
Smith, J.H.G. 1978. Management of Douglas-fir and other forest types in the Vancouver Public Sustained Yield Unit. In *Forest management in Canada*, vol. 11. Ottawa: Environment Canada.
Von Thünen, J.H. 1826. *The isolated state*. Trans. and ed. Peter Hall. London: Pergamon Press.

Foster, William Trufant (1879–1950)

Robert W. Dimand

Keywords

American Economic Association; Catchings, W.; Foster, W.; Great Depression; Growth, models of; Harrod–Domar growth theory; Hoover Plan; Monetary policy rules; Paradox of thrift; Pollak Foundation; Recessions; Underconsumption

JEL Classifications

B31

The educator and heterodox monetary economist William Trufant Foster was born in Boston, Massachusetts, on 18 January 1879, and died in Winter Park, Florida, on 18 October 1950. After his father's early death, Foster worked his way through high school and Harvard University, graduating first in his class in 1901. After teaching at Bates College in Lewiston, Maine, he returned to Harvard to take an A.M. in English in 1904, followed by a Ph.D. from Teachers College of Columbia University. His exceptional success as a teacher of rhetoric and a textbook author, and the vision of an

'ideal college' presented in his doctoral dissertation (published in 1911), led to his remarkably early promotion from instructor to full professor at Bowdoin College in Brunswick, Maine, in 1905, and his appointment at the first president of Reed College in Portland, Oregon, in 1910. Foster served as an inspector with the American Red Cross in France after US entry into the First World War. Health problems from overwork, together with controversy over his pacifism, led Foster to resign from Reed College in December 1919. He then became director of the Pollak Foundation for Economic Research, founded in Newton, Massachusetts, by his Harvard classmate Waddill Catchings, an investment banker.

The Pollak Foundation was a vehicle for expounding the heterodox monetary theories of Foster and Catchings, and, through Houghton Mifflin, published their books on *Money* (1923), *Profits* (1925), and *Business without a Buyer* (1927a). They held that recessions, such as that of 1920–1, happen because a monetary economy does not automatically generate enough consumption to buy potential output. Saving, which enriches the individual saver, contributes to recessions both by reducing consumption and, through investment, by adding to the potential output to be purchased. Because of this paradox of thrift and their support for counter-cyclical public works, Foster and Catchings had been considered as possible forerunners of Keynesian macroeconomics, while their emphasis on a steadily increasing rate of investment as a prerequisite for stable growth has been related to later Harrod–Domar growth theory (see Gleason 1959; Carlson 1962). Their support for the proposal by Carl Snyder of the Federal Reserve Bank of New York that the volume of currency and credit be increased by a steady four per cent a year has been suggested as an anticipation of monetarist policy rules (Tavlas 1976). In late 1928, with President-elect Hoover's endorsement and with Foster as his expert witness, Governor Ralph Brewster of Maine submitted to the annual governors' conference a plan for standby credit authorization for $3 billion of federal, state and local public works to be undertaken once a federal board certified the imminence of a recession (Dorfman 1959; Barber 1985). Although Foster and Catchings promoted this as the 'Hoover Plan', once the Depression hit President Hoover felt that budget deficits precluded such large-scale counter-cyclical public works.

The Pollak Foundation offered a $5,000 prize for the best adverse criticism of Foster and Catchings's *Profits,* with the competition judged by the two most recent presidents of the American Economic Association, Wesley Mitchell and Allyn Young, and by Owen Young of General Electric. The competition attracted 431 submissions, and the four winning essays were published with a reply by Foster and Catchings as *Pollak Prize Essays* (1927b). Also in 1927, the magazine *World's Work* offered a $1,000 prize for the best essay on a series of articles by Foster and Catchings in the magazine. These prizes brought Foster and Catchings considerable professional attention, as did the Pollak Foundation's publication of substantial studies of index numbers by Irving Fisher and of real wages by Paul Douglas (who had been Foster's student at Bowdoin and junior colleague at Reed). Foster and Catchings also found a more popular audience: *The Road to Plenty* (1928), presented as a conversation aboard a train, sold 58,000 copies, while *Progress and Plenty* (1930) reprinted 206 of their 400 two-minute talks on economic problems distributed by the McClure Newspaper Syndicate in 1929 and 1930.

Financial difficulties forced Catchings to withdraw from active participation in the Pollak Foundation during the Depression. Foster continued to direct the foundation, and for three years in the 1930s wrote a syndicated daily newspaper column on economics for the layperson. He served on the Consumers Advisory Board of the National Recovery Administration from 1933 to 1935 (recommended by Paul Douglas) and was an economic adviser at the International Labor Conference in Geneva in 1938.

See Also

- ▶ Catchings, Waddill (1879–1967)
- ▶ Monetary Cranks
- ▶ Underconsumptionism

Selected Works

1911. *Administration of the college curriculum*. New York: Columbia University Press.
1923. (With W. Catchings.) *Money.* Boston: Houghton Mifflin.
1925. (With W. Catchings.) *Profits.* Boston: Houghton Mifflin.
1927a. (With W. Catchings.) *Business without a buyer.* Boston: Houghton Mifflin.
1927b. (With W. Catchings.) *Pollak prize essays.* Newton: Pollak Foundation for Economic Research.
1928. (With W. Catchings.) *The road to plenty.* Boston: Houghton Mifflin
1930. (With W. Catchings.) *Progress and plenty.* Boston: Houghton Mifflin.

Bibliography

Barber, W. 1985. *From new era to new deal: Herbert Hoover, the economists, and American economic policy, 1921–1933*. Cambridge: Cambridge University Press.
Carlson, J. 1962. Foster and Catchings: A mathematical reappraisal. *Journal of Political Economy* 70: 400–402.
Dorfman, J. 1959. *The economic mind in American civilization. Volumes 4 and 5: 1918–1933*. New York: Viking.
Gleason, A. 1959. Foster and Catchings: A reappraisal. *Journal of Political Economy* 67: 156–172.
Tavlas, G. 1976. Some further observations on the monetary economics of Chicagoans and non-Chicagoans. *Southern Economic Journal* 42: 685–692, with comment by J. Davis and reply by Tavlas, 45 (1979), 919–931.

Fourier, FranÇois Marie Charles (1772–1837)

J. Wolff

According to Fourier, the poverty which accompanies the 'colossal' advance of industry constitutes the main cause of social disorders. Those responsible are the tradesmen; competition has resulted in the creation of a mercantile feudal system.

To set up a new social order, we must use the passions as nature gave them to us. Man is guided by his quest for pleasure, and his passions are always the same, the principal ones being the desire for luxury, the desire to adhere to a group, and that of forming part of a 'series' or a work or play group.

These groupings of individuals must be such that they allow psychological differences to complement each other. Individuals should live together in a 'phalanstery', where, in giving themselves over entirely to their passions, they will form a harmonious and pacific social order. The association will encourage emulation and the disappearance of rivalry. Work will be almost infinitely divided up, its duration will be short, and everyone will be free to choose the work of his choice. Associated property will consist of property brought by the participants who do not keep goods for themselves. Manual work, capital and 'talent' will be remunerated.

The application of Fourier's system has been attempted several times, and has generally ended in failure. However, the experiment led by J.B. Godin at Guise in France can be considered half successful as it was followed up after his death and lasted until 1969.

Fourier was essentially concerned with psychology and social psychology, and can be seen as the precursor of studies conducted from the 1920s onwards on the ways in which work groups function.

Selected Works

1808. Théorie des quatre mouvements et des destinées générales. 1822. Traité de l'association domestique agricole.
1829. Le nouveau monde industriel et sociétaire. 1835–6. La fausse industrie.
The complete works of Fourier were republished in eleven volumes by Editions Anthropos, Paris, 1966, a reprint of the third edition of 1846.

Bibliography

Bourgin, H. 1905. *Fourier*. Paris.
Lehouck, E. 1966. *Fourier aujourd'hui*. Paris: Denoel.
Pinloche, A. 1933. *Fourier et le socialisme*. Paris: F. Alcan.

Foxwell, Herbert Somerton (1849–1936)

Gerard M. Koot

Born at Shepton Mallet, Somerset, Foxwell received his early education at home and then at schools in Taunton. He matriculated into the University of London in 1866 and received his BA in 1868. He entered St John's College, Cambridge, in 1868. After being placed Senior in the Moral Science Tripos in 1870, he won the Whewell Scholarship in International Law in 1872. He was elected a Fellow of the College in 1874. Under the old Statutes he was forced to vacate his fellowship when he married in 1898, but was able to resume it in 1905 and retained it until his death in 1936. From 1875 until 1905 he served as College Lecturer. At first he taught the whole area of the Moral Sciences, but while Marshall was at Bristol from 1877 until 1884, Foxwell taught courses in economics. When Marshall became Professor of Political Economy at Cambridge in 1885, he quickly overshadowed Foxwell both at St John's and in Cambridge economics generally. Appointed as a University Extension Lecturer in 1874, Foxwell taught widely in the North of England, claiming that it had brought him into close contact 'with the actual conditions of practical life'. In 1876 he became a Lecturer at University College, London, and in 1881 he succeeded W.S. Jevons as its Professor of Political Economy. Despite his frequent travels to London, Foxwell remained firmly committed to Cambridge life. After his retirement from active lecturing at St John's, he served as its Director of Economic Studies until his death. In private life, he lived a few doors from J.N. Keynes in Harvey Road and was well known in Cambridge economic circles.

Foxwell's primary interests were in bibliography, banking and money, and economic history. As a book collector, he assembled several large libraries of economic literature, especially for the period 1740–1848, which became the basis for the Goldsmiths' collection in London and the Kress Library of Business and Economics at Harvard. Along with Jevons, whom he knew well, Foxwell was a severe critic of Ricardo. Indeed, one of the aims of his book-collecting was to demonstrate that England had produced other economic traditions than that of Ricardo. Foxwell held that Ricardo's use of deduction had been excessive and that it had produced a tradition of socialist thought. Foxwell's own conservative position is demonstrated in his historical introduction to Anton Menger's *Right to the Whole Produce of Labour* (1899).

Foxwell attributed his interest in the instability of capitalism to both Jevons and Arnold Toynbee. His major original work in economics, *Irregularity of Employment and Fluctuations of Employment* (1886) held that free competition had produced both wealth and poverty for the workers. Poverty, he argued, was primarily a result of the irregularity of employment due to an unstable level of prices, as well as the persistence of a customary low level of wages for many groups in society. Fearful of social revolution, and building on the work of J.S. Mill, he fashioned a counter-revolutionary programme of state intervention, cooperative schemes, profit sharing, and the benefits which could be derived from regulated monopolies. He especially promoted a system of counter-cyclical state expenditures in such areas as housing, health, education and public works. A strong advocate of bimetallism, he also called for the adoption of a managed system of international payments.

In 1887, Foxwell published his influential 'Economic Movement in England' in the *Quarterly Journal of Economics*, in which he sympathetically chronicled the rise of an ethical and historical economics in England which he claimed as a superior guide to the formulation of public policy than the dominant Ricardian tradition. Subordinate to Marshall at Cambridge, Foxwell increasingly allied himself with the historical criticism of Marshall's economics. In 1903 he even joined the historical economists' attack on free trade. From 1895 he lectured at the London School of Economics and Political Science, which had been established in part in conscious

opposition to Marshall's vision of economics, and in 1907 was named Professor of Political Economy at the University of London. His implacable hostility to Ricardo, whom Marshall vigorously defended, his increasing attention to economic history and bibliographical work, his inability to contribute to the development of neoclassical economic theory, his bimetallism, and his difficult personality, made him unacceptable to Marshall as his successor in 1908. Instead, the appointment when to the able and more suitable theorist A.C. Pigou. It was not until 1929, when he was elected President of the Royal Economic Society, that Foxwell reconciled himself to having been passed over by Marshall at Cambridge.

See Also

- ▶ Economics Libraries and Documentation
- ▶ Marshall, Alfred (1842–1924)

Selected Works

1886. Irregularity of employment and fluctuations of employment. Lecture VI in the *claims of labour*. Edinburgh: Cooperative Printing Company.

1887. The economic movement in England. *Quarterly Journal of Economics* 2: 84–103.

1899. Introduction and Bibliography to Anton Menger. In *The right to the whole produce of labor*. London. Reprinted, New York: Augustus M. Kelley, 1968.

References

Coase, R.H. 1972. The appointment of Pigou as Marshall's successor. *Journal of Law and Economics* 15: 473–486.

Coats, A.W. 1972. The appointment of Pigou as Marshall's successor: A comment. *Journal of Law and Economics* 15: 487–495.

Foxwell, A.G.D. 1939. Herbert Somerton Foxwell: A portrait. In *Kress library of business and economics publication,* No. 1. Boston: Harvard Graduate School of Business Administration, 3–30.

Keynes, J.M. 1936. H.S. Foxwell. *Economic Journal* 46: 589–614. Reprinted in *The collected works of John Maynard Keynes* Vol. X, *essays in biography.* London: Macmillan, 1972.

Koot, G.M. 1977. H.S. Foxwell and English historical economics. *Journal of Economic* 11(2): 561–586.

Fractals

Laurent E. Calvet

Abstract

Fractals have become increasingly useful tools for the statistical modelling of financial prices. While early research assumed invariance of the return density with the time horizon, new processes have recently been developed to capture nonlinear changes in return dynamics across frequencies. The Markov-switching multifractal (MSM) is a parsimonious stochastic volatility model containing arbitrarily many shocks of heterogeneous durations. MSM captures the outliers, volatility persistence and power variation of financial series, while permitting maximum likelihood estimation and analytical multi-step forecasting. MSM compares favourably with standard volatility models such as GARCH(1,1) both in-and out-of-sample.

Keywords

Bayesian filtering; Brownian motion; Continuous time valuation; Fractals; Fractional Brownian motion; Lévy-stable processes; Long memory models; Mandelbrot, B; Markov-switching multifractal (MSM); Maximum likelihood; Multifractal model of asset returns (MMAR); Multifractality; Regime switching; Self-similarity processes

JEL Classification

B23; C22; C53; G1; D85

The word 'fractal' was coined by the French mathematician Benoît Mandelbrot (1982) to

characterize a wide class of highly irregular scale-invariant objects. It originates from the Latin adjective *fractus*, meaning 'broken' or 'fragmented'. The defining characteristic of fractals is that their degree of irregularity remains the same at all scales. This invariance permits parsimonious modelling of complex objects, and has been useful for analysing a wide variety of natural phenomena. The entry reviews the use of fractals in economics and finance, and more specifically their application in the statistical modelling of asset returns, which has been a remarkably active field since the early 1960s.

Consider the price $P(t)$ of a financial asset, such as a stock or a currency, and let $p(t)$ denote its logarithm. The process $p(t)$ is said to be *self-similar* if there exists a constant $H > 0$ such that for every set of instants $t_1 \leq \ldots \leq t_k$ and for every $\lambda > 0$, the vector $\{p(\lambda t_1), \ldots, p(\lambda t_k)\}$ has the same distribution as $\lambda^H\{p(t_1), \ldots, p(t_k)\}$,

that is,

$$\{p(\lambda t_1), \ldots, p(\lambda t_k)\} \underline{d} \{\lambda^H p(t_1), \ldots, \lambda^H p(t_k)\}. \quad (1)$$

The constant H is called the *self-similarity* index.

Three classes of self-similar processes have been widely used in finance: the Brownian motion, Lévy-stable processes and the fractional Brownian motion, which are successively discussed. The Brownian motion (Bachelier 1900), with self-similarity index $H = 1/2$, pervades modern financial theory and notably the Black–Merton–Scholes approach to continuous time valuation. Its lasting success arises from several appealing properties, including tractability and consistency with the financial concepts of no-arbitrage and market efficiency.

The stable processes of Paul Lévy (1924) are characterized by thicker tails than the Brownian motion. They are thus more likely to accommodate the outliers exhibited by financial series, as was pointed out by Mandelbrot in a series of seminal papers (for example, 1963). The increments of Lévy-stable processes are stationary and have stable distributions, where stability refers to invariance under linear combinations (see Samorodnitsky and Taqqu 1994). Tails are Paretian:

$$\mathbb{P}\{p(\Delta t) > x\} \sim cx^{-\alpha} \text{ as } x \to +\infty,$$

with index $\alpha = 1/H \in (0;2)$. The variance of a Lévy-stable process is infinite, which is at odds with both empirical evidence and mean-variance asset pricing. Furthermore, stable processes have independent increments and thus cannot account for volatility clustering.

The fractional Brownian motion (Kolmogorov 1940; Mandelbrot 1965; Mandelbrot and Van Ness 1968) with $H > 1/2$ is a self-similar process with strongly dependent returns. Increments are stationary, correlated, and normally distributed. Their autocorrelation declines at the hyperbolic rate

$$Cov[r(t); r(t+n)] \sim c\,(2H-1)\,n^{2H-2} \text{ as } n \to \infty$$

where $r(t) = p(t) - p(t - \Delta t)$ denotes the return on a time interval of fixed length Δt. Hyperbolic autocorrelation is the defining property of long-memory processes, whose use in economics was advanced by the discrete-time fractional integration approach of Granger and Joyeux (1980). While research on long memory has generally been very fruitful in economics (see Baillie 1996, for a review), the fractional Brownian motion rarely represents a practical model of asset prices. Specifically, long memory in returns is both empirically inaccurate in most markets (Lo 1991) and inconsistent with arbitrage-pricing in continuous time (Maheswaran and Sims 1993). There is, however, abundant evidence of long memory in the *volatility* of returns (for example, Dacorogna et al. 1993; Ding et al. 1993).

In all the above self-similar processes, returns observed at various frequencies have identical distributions up to a scalar renormalization:

$$p(t + \lambda\Delta t) - p(t) \underline{d} \lambda^H p(\Delta t).$$

Most financial series, however, are not exactly self-similar, but have thicker tails and are more peaked in the bell at shorter horizons. This observation is consistent with the economic intuition that high-frequency returns are either large if new information has arrived, or close to zero otherwise. Thus, self-similar processes do not capture

in a single model the most salient features of asset returns.

A partial solution to these difficulties is provided by the multifractal model of asset returns (MMAR; Calvet et al. 1997; Calvet and Fisher 2002a). This approach builds on multifractal measures (Mandelbrot 1974), which are constructed by the iterative random reallocation of mass within a time interval. The MMAR extends multifractals from measures to diffusions. The asset price is specified by compounding a Brownian motion with an independent random time-deformation:

$$p(t) = B[\theta(t)],$$

where θ is the cumulative distribution of a multifractal measure $\theta(t) = \mu[0,t]$ Returns are uncorrelated and the price p is a martingale in MMAR, which precludes arbitrage. The time deformation induces sharp outliers in returns and long memory in volatility. The MMAR also captures nonlinear changes in the return density with the time horizon (Lux 2001).

The price p inherits highly heterogeneous time-variations from the multifractal measure. Its sample paths are continuous but can be more irregular than a Brownian motion at some instants. Specifically, the local variability of a sample path at a given date t is characterized by the local Hölder exponent $\alpha(t)$, which heuristically satisfies

$$|p(t+dt) - p(t)| \sim c_t (dt)^{\alpha(t)} \text{ as } dt \to 0.$$

Traditional jump diffusions impose that $\alpha(t)$ be equal to 0 at points of discontinuity, and to 1/2 otherwise. In a multifractal process, however, the exponent $\alpha(t)$ takes a *continuum* of values in any time interval.

Asset returns at different frequencies satisfy the moment-scaling rule:

$$\mathbb{E}[|p(\Delta t)|^q] = c_q (\Delta t)^{\tau(q)+1},$$

which holds for every (finite) moment q and time interval Δt. These moment restrictions represent the basis of estimation and testing (Calvet et al. 1997; Calvet and Fisher 2002a, b; Lux 2004). The MMAR provides a well-defined stochastic framework for the analysis of moment-scaling, which has generated extensive interest in econophysics (for example, LeBaron 2001). The multifractal model is also related to recent econometric research on power variation, which interprets return moments at various frequencies in the context of traditional jump-diffusions (for examples, Andersen et al. 2001; Barndorff-Nielsen and Shephard 2004).

Despite its appealing properties, the MMAR is unwieldy for econometric applications because of two features of the underlying measure: (*a*) the recursive reallocation of mass on an entire time-interval does not fit well with standard time series tools; and (*b*) the limiting measure contains a residual grid of instants that makes it non-stationary.

The Markov-switching multifractal (MSM) resolves these difficulties by constructing a fully stationary volatility process that evolves stochastically through time (Calvet and Fisher 2001, 2004). MSM builds a bridge between multifractality and regime-switching, which permits the application of Bayesian filtering and maximum likelihood estimation to a multifractal process. Volatility is driven by the first-order Markov state vector $M_t = \left(M_{1,t}; M_{2,t}; \ldots; M_{\bar{k},t}\right) \in \mathbb{R}_+^{\bar{k}}$, whose components have unit mean and heterogeneous persistence levels. In discrete time, returns are specified as

$$r_t = \sigma\left(M_{1,t} M_{2,t} \ldots M_{\bar{k},t}\right)^{1/2} \varepsilon_t, \qquad (2)$$

where σ is a positive constant and $\{\varepsilon_t\}$ are independent standard Gaussians. Volatility components follow independent Markov processes that are identical except for time scale. Given the volatility state M_t, the next-period multiplier $M_{k;t+1}$ is drawn from a fixed distribution M with probability γ_k, and is otherwise left unchanged.

Components differ in their transition probabilities γ_k but not in their marginal distribution M. The transition probabilities are tightly specified by $\gamma_k = 1 - (1-\gamma_1)^{\left(b^{k-1}\right)}$, which is approximately geometric at low frequency: $\gamma_k \sim \gamma_1 b^{k-1}$. In empirical applications, a unique scalar m_0

typically determines the distribution M. The return process (2) is then specified by the four parameters (m_0, σ, b, γ^1). Since the number of frequencies k can be arbitrarily large, MSM provides a tight specification of a high-dimensional state space. The approach conveniently extends to continuous time (Calvet and Fisher 2001) or a multivariate setting (Calvet et al. 2006).

When M has a discrete distribution, the state space is finite and MSM defines a stochastic volatility model with a closed-form likelihood. It then bypasses the estimation problems of traditional stochastic volatility settings based on smooth autoregressive transitions. On the other hand when M has a continuous (for example, lognormal) distribution, estimation can proceed by simulated method of moments (Calvet and Fisher 2002b), generalized method of moments (Lux 2004), or simulated likelihood via a particle filter (Calvet et al. 2006).

MSM tends to substantially outperform traditional models both in and out of sample. Calvet and Fisher (2004) thus report considerable gains in exchange rate volatility forecasts at horizons of 10–50 days as compared with GARCH-type processes. Lux (2004) obtains similar results with lognormal MSM using linear predictions. Furthermore, bivariate MSM compares favourably with multivariate GARCH under criteria such as the likelihood function, integral transforms and value-at-risk (Calvet et al. 2006).

The integration of multifrequency models into asset pricing is now at the forefront of current research. Calvet and Fisher (2005a) thus introduce a parsimonious equilibrium set-up in which regime shifts of heterogeneous durations affect the volatility of dividend news. The resulting return process is endogenously skewed and has significantly higher likelihood than the classic Campbell and Hentschel (1992) specification. Calvet and Fisher (2005b) similarly illustrate the potential of MSM for building parsimonious multifrequency jump-diffusions.

See Also

► Regime Switching Models

Bibliography

Andersen, T., T. Bollerslev, F. Diebold, and P. Labys. 2001. The distribution of realized exchange rate volatility. *Journal of the American Statistical Association* 96: 42–55.

Bachelier, L. 1900. Théorie de la spéculation. *Annales Scientifiques de l'Ecole Normale Supérieure* 17: 21–86.

Baillie, R. 1996. Long memory processes and fractional integration in econometrics. *Journal of Econometrics* 73: 5–59.

Barndorff-Nielsen, O., and N. Shephard. 2004. Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics* 2: 1–37.

Calvet, L., and A. Fisher. 2002a. Multifractality in asset returns: Theory and evidence. *Review of Economics and Statistics* 84, 381–406.

Calvet, L., and A. Fisher. 2002b. Regime-switching and the estimation of multifractal processes. Working paper. Harvard University and University of British Columbia.

Calvet, L., and A. Fisher. 2005a. Multifrequency news and stock returns. Working paper no. 11441. Cambridge, MA: NBER.

Calvet, L., and A. Fisher. 2005b. Multifrequency jump diffusions: An equilibrium approach. Working paper. HEC School of Management and University of British Columbia.

Calvet, L., and A. Fisher. 2001. Forecasting multifractal volatility. *Journal of Econometrics* 105: 27–58.

Calvet, L., and A. Fisher. 2004. How to forecast long-run volatility: Regime-switching and the estimation of multifractal processes. *Journal of Financial Econometrics* 2: 49–83.

Calvet, L., A. Fisher, and B. Mandelbrot. 1997. *A multifractal model of asset returns*, Discussion papers 1164–1166. New Haven: Cowles Foundation, Yale University.

Calvet, L., A. Fisher, and S. Thompson. 2006. Volatility comovement: A multifrequency approach. *Journal of Econometrics* 131: 179–215.

Campbell, J., and L. Hentschel. 1992. No news is good news: An asymmetric model of changing volatility in stock returns. *Journal of Financial Economics* 31: 281–318.

Dacorogna, M., U. Müller, R. Nagler, R. Olsen, and O. Pictet. 1993. A geographical model for the daily and weekly seasonal volatility in the foreign exchange market. *Journal of International Money and Finance* 12: 413–438.

Ding, Z., C. Granger, and R. Engle. 1993. A long memory property of stock returns and a new model. *Journal of Empirical Finance* 1: 83–106.

Granger, C., and R. Joyeux. 1980. An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis* 1: 15–29.

Kolmogorov, A. 1940. Wienersche Spiralen und einige andere interessante Kurven im Hilbertschen raum. *Comptes Rendus de l'Académie des Sciences de l'URSS* 26: 115–118.

F

LeBaron, B. 2001. Stochastic volatility as a simple generator of apparent financial power laws and long memory. *Quantitative Finance* 1: 621–631.
Lévy, P. 1924. Théorie des erreurs: la loi de Gauss et les lois exceptionnelles. *Bulletin de la Société Mathématique de France* 52: 49–85.
Lo, A. 1991. Long memory in stock market prices. *Econometrica* 59: 1279–1313.
Lux, T. 2001. Turbulence in financial markets: The surprising explanatory power of simple cascade models. *Quantitative Finance* 1: 632–640.
Lux, T. 2004. The Markov-switching multifractal model of asset returns: GMM estimation and linear forecasting of volatility. Working paper. Kiel University.
Maheswaran, S., and C. Sims. 1993. Empirical implications of arbitrage-free asset markets. In *Models, methods and applications of econometrics*, ed. P. Phillips. Oxford: Blackwell.
Mandelbrot, B. 1963. The variation of certain speculative prices. *Journal of Business* 36: 394–419.
Mandelbrot, B. 1965. Une classe de processus stochastiques homothétiques à soi. *Comptes Rendus de l'Académie des Sciences de Paris* 260: 3274–3277.
Mandelbrot, B. 1974. Intermittent turbulence in self-similar cascades: Divergence of high moments and dimension of the carrier. *Journal of Fluid Mechanics* 62: 31–58.
Mandelbrot, B. 1982. *The fractal geometry of nature*. New York: Freeman.
Mandelbrot, B., and J. Van Ness. 1968. Fractional Brownian motion, fractional noises and applications. *SIAM Review* 10: 422–437.
Samorodnitsky, G., and M. Taqqu. 1994. *Stable non-gaussian random processes*. New York: Chapman and Hall.

France, Economics in (After 1870)

Alain Béraud and Philippe Steiner

Abstract

After 1870, classical liberals gradually lost their influence. Political economy began to be taught in university faculties of law, and also in some of the engineering schools. This laid the foundations for a long-standing divide between two groups of economists. Professors of political economy in the law faculties often inclined to an institutionalist approach, and opposed the mathematical approach to political economy that economic engineers and some mathematicians adopted. This antagonism abated after the Second World War as French economists strengthened their relations with foreign colleagues.

Keywords

Aftalion, A; Allais, M; Aupetit, A; Austrian School; Bachelier, L; Balasko, Y; Barrère, A; Baudin, L; Benassy, J. -P; Bertrand, J; Birth rate; Boiteux, M; Borel, E; Boyer, R; Braudel, F; Brownian motion; Budgetary equilibrium; Cartan, H; Chamberlin, E; Classical liberalism; Colson, C; Condillac, E; Consumer price index; Contract theory; Convention School; Correlation analysis; Cournot, A; Courtin, R; Creative destruction; Currency School; D'Aspremont, C; Debreu, G; Decisions under uncertainty; Demand for money; Differentiability; Disequilibrium theory; Divisia, F; Dos Santos, R; Drèze, J; Econometrics; Economic growth; Efficient markets; Excess labour supply; Existence of equilibrium; Expectations; Family planning; General equilibrium; Gérard-Varet, L. -A; German Historical School; Gide, C; Golden rule; Grandmont, J.-M; Gruson, C; Guesnerie, R; Hicks, J; Identification; Indicative planning; Inflation; Innovation; Institut de Sciences Économiques Appliquées; Institutionalism; Intertemporal choice; IS–LM model; Jevons, W; Juglar, C; Keynes. J. M; Labrousse, E; Lacaillon, J; Laffont, J. -J; Landry, A; Lenoir, M; Leroy- Beaulieu, P; Lescure, J; Lindahl, E; Local equilibrium; Malinvaud, E; March, L; Marchal, A; Marchal, J; Marjolin, R; Marxism; Massé, P; Mathematical economics; Method of moments; Minimax theorem; Modigliani, F; Molinari, G. de; Monetary theory of crises; Money supply; Mortality; Nogaro, B; Oligopolistic competition; Orléan, A; Overlapping generations model; Partial equilibrium; Pearson, K; Perroux, F; Probability; Purchasing power parity; Quantity theory of money; Regulation; Regulation School; Risk; Rist, C; Roy, R; Rueff, J; Saillard, Y; Say, J. -B; Say's Law; Shadow prices; Simiand, F; System of curves; Tâtonnement; Tirole, J; Transaction costs; Transfer problem; Unemployment; Villey, D;

Von Neumann, J; Wage determination; Walras, L; Younès, Y

JEL Classifications
B1

The publication of Léon Walras's *Éléments d'économie politique pure* in 1874 marks an important turning point in the history of economic analysis. But for many years his ideas remained misunderstood. Recognition of the importance of his work on the part of French economists followed a lengthy and difficult period, in which the publication of Maurice Allais's *À la recherche d'une discipline économique, l'économie pure* in 1943 marks a vital stage. Allais introduced the analysis of risk and intertemporal choice to the theory of general equilibrium and in this way posed new questions to which Gérard Debreu, Marcel Boiteux, Edmond Malinvaud and many others would respond. Nonetheless, many French economists had considerable reservations about the theory of general equilibrium. They favoured an emphasis upon the role of institutions, and the need to integrate the various elements of the social sciences – economics, sociology and history – if economic phenomena were to be understood.

From 1870 to 1943

In the years after 1870 the domination of the liberal school was increasingly questioned, and this was largely the consequence of institutional developments (Le Van-Lemesle 2004). The teaching of political economy was introduced into the faculties of law in 1877, but the professors in law in charge of this teaching progressively became scientifically independent. In 1887 they founded the *Revue d'Économie Politique* so that the new political economy might be more widely diffused, and this quickly became far more influential than the liberal *Journal des Économistes.*

Classics Liberals and Institutionalists

The best known of the last classical liberals, Gustave de Molinari and Paul Leroy- Beaulieu, sought to defend very different positions. Liberals had maintained that the state should limit itself to the provision of individual security but de Molinari (*L'évolution politique et la révolution,* 1884) argued that it was necessary to go much further. All branches of production, including the judiciary, the police and defence, should be freed from state control. If a need for security exists and if the state does not foresee it, then this need will be met by private initiative, and so much the better. Leroy-Beaulieu did not challenge the principle that a state had its prerogatives and that it would exercise them. However, while Molinari defended the classical theory of distribution, Leroy-Beaulieu (*Essai sur la répartition des richesses,* 1881) thought it necessary to abandon this theory. The consequences which it foresaw – a fall in the rate of profit, an increase in the rate of rents, and a reduction of wages to subsistence levels – were refuted by factual evidences: wage rates were increasing, and rents were diminishing in proportion. Institutionally Leroy-Beaulieu belonged to the group of older classical liberals, but he abandoned the propositions basic to this school.

Charles Gide occupied a leading place among the professors of the law faculties. He was a staunch eclectic, which led him to reject extreme theses in favour of an intermediate synthesis. In studying prices and distribution he made use of ideas borrowed from Jevons and Walras, but played down their contribution. If Jevons's analysis of value was ingenious, it was nonetheless not new; Condillac had long before made clear that the utility of an object determined its value. Gide was somewhat reluctant to make use of the notion of marginal productivity, since he did not consider that the distribution of revenues to be solely determined by economic factors, and he argued (*Principes d'économie politique,* 1901) that social relations among agents also played a part.

Adolphe Landry and François Simiand were part of a very small group of philosophers educated at the École Normale who chose to become economists. In his *Révolution démographique* (1934) Landry distinguished three types of regulation as of importance to the study of demographic development. First, under the *ancien*

régime, parents did not concern themselves with the consequences of the birth of children. Mortality played the principal role in regulating the population. Second, during the transitional phase, men and women chose their age of marriage so that they might maintain the standard of living to which they had become accustomed, and there was no voluntary birth control in marriage. Third, in modern times, on the contrary, the timing and number of births had become a matter of choice. Landry used this argument to persuade parliament to vote through, in 1932, 1939 and 1946, the three laws which determine the allocations of family support: for if the birth rate is the product of choice, then one can hope to end demographic decline with the aid of a system of financial incentives.

French positive economics developed with the work of François Simiand (*La méthode positive en science économique,* 1912). He rejected both the approach of the German Historical School as well as what he termed 'orthodox' economics, referring in this way to French liberals, the Austrian School and mathematical economics. The German Historical School, he suggested, lacked principles and had produced nothing but an empty accumulation of knowledge. 'Orthodox' economists constructed theories that were poorly founded, since they drew upon incomplete or implicit observations. Simiand, by contrast, made use of long statistical series, analysing them in terms of models that described the behaviour of social groups. He applied this method to the study of the development of wages and prices in his major works of the 1930s (*Recherches anciennes et nouvelles sur le mouvement des prix du 16ème au 19ème siècle,* 1932, et *Le salaire, l'évolution sociale et la monnaie,* 1932). In these works he argued that variations in the money supply drove the cycle and that cyclical fluctuation was a necessary part of economic progress. This approach influenced Ernest Labrousse (*Esquisse du mouvement des prix et des revenus au 18ème siècle,* 1933) who, on the basis of meticulously constructed statistical series, put forward a simple theory of the crisis of the *ancien régime* as engendered by the agricultural cycle: bad harvests brought about a rise in the price of wheat, consumers spent an increasing proportion of their revenues on agricultural goods and so the crisis was transmitted to industry.

Albert Aftalion (*Les crises périodiques de surproduction,* 1913) and Jean Lescure (*Des crises générales et périodiques de surproduction,* 1906) took their inspiration from Say and their analysis of crises from Juglar. They retained Say's Law of Markets. From Juglar they drew three lessons. Their analysis rested upon study of empirical data. They used price movements to determine the phases of the cycle. The crisis was defined as the point at which prices ceased rising, inevitably followed by a fall in prices – it was only one phase of the cycle. But whereas Juglar put forward a monetary theory of crises, Aftalion and Lescure proposed a real theory. At the bottom of a recession production had difficulty satisfying needs. The marginal utility of consumer goods and their prices would thus rise. To meet this demand, machinery is needed. The price of machinery rises and in turn stimulates production. However, when the new production goods come into service consumer goods become overabundant. Their final utility and value collapse and this has repercussions for the price of machinery. The crisis becomes almost, or entirely, general. Lescure placed the emphasis on the role of profits and on the interdependence between activities. At the end of an expansionary phase, costs rise faster than prices and new enterprises that have paid a high price for their means of production face losses. Their insolvency brings about the crisis, which spreads from one branch to another. The crisis is not general, but generalized.

Over a lengthy period, French economists had criticized the version of the quantity theory of money advocated by partisans of the Currency School, and this continued after 1870. Bertrand Nogaro (*Contribution à une théorie réaliste de la monnaie,* 1906) noted that money was the object neither of demand nor supply; the general price level is not determined, as the quantity theory supposed, by the relation between the money stock and desired cash holdings, but by global demand for goods, or as argued by Aftalion (*Monnaie, prix et change,* 1927), by the relationship between monetary revenue and the volume of

production. The consequences of a variation in the stock of money depended for its effect upon the demand and supply of goods, and hence on the way that it is introduced into the system. Nogaro and Aftalion rejected the idea that variations in the price of goods explained variations in the exchange rate. The direction of causality was not necessarily from prices to exchange rates. The current exchange rate depended upon the expected future rate and, since it affected producer costs and agents' revenues, domestic prices are determined by psychological factors.

Walras, the Mathematicians and the Statisticians

For many years both mathematicians and engineers had reservations about the idea of general equilibrium. They considered partial equilibrium quite adequate for the study of most problems. Walras's use of mathematics seemed quite superfluous. Even when the importance of Walras's work gradually became more generally accepted, his successors remained critical of his methodology. Instead they shared Pareto's view that the criterion of a theory's truth lies in its correspondence to reality. They did not attempt to resolve the theoretical difficulties presented by the Walrasian construct. Instead, they were interested in understanding the instruments which permitted the analysis of facts while using economic theory. The procedure followed by Albert Aupetit, the leading disciple of Walras, is quite typical. His dissertation, *Essai sur la théorie génerale de la monnaie* (1901), presents itself both as a development of Walrasian monetary theory and as verification of its empirical relevance.

The tradition of engineer–economists continued with Clément Colson. His works (*Cours d'économie politique,* 1901–7) drew more on Dupuit's analysis than on Walras's, but he encouraged François Divisia, René Roy and Jacques Rueff to study Walrasian theory since he was aware of the importance of the interdependence of markets. It was not possible to study the determination of wages independently of that of the rate of interest. Since labour and capital are substitutes, the proportions in which they should be employed depended both upon the wage rates and interest rates. Here one can see at work the fundamental idea that had driven Walras to use mathematics and make use of models of general equilibrium.

Divisia's analysis of monetary phenomena illustrates this connection of theory to empirical research. It had sometimes been thought that the quantity equation implies that prices vary with the quantity of money. Divisia rejected this idea, arguing that the transactions equation is an identity. Appealing to statistical observation for verification is an absurdity, but it does allow the definition of what should be an indicator of prices. Weights are quantities of goods and services exchanged, not quantities produced or consumed. Divisia (*L'indice monétaire et la théorie de la monnaie,* 1925–6) explained that it is not possible to set these weights; the index should be a chain index. In order to determine the value of money in 1900 relatively to its value in 1800, it is not enough to know the quantities of goods and services bought in 1800 and 1900, all the intermediate values should also be known. René Roy followed the same line of argument. He introduced (*De l'utilité, contribution à une théorie des choix,* 1942) the idea of the indirect utility function to demonstrate that the consumer price index is the number by which primary prices have to be multiplied to render the satisfaction of an individual (under the assumption of constant monetary income) equal to his satisfaction at current prices.

Even while invoking Walras, Rueff appeared above all to be the defender of classical arguments against attack by Institutionalists and by Keynes. Contrary to Nogaro, he argued (*Théorie des phénomènes monétaires,* 1926) that price variations are determined by effective holdings of cash relative to desired holdings. He based his arguments on a reformulation of the theory of purchasing power parity in dealing with the problem of transfers. Contrary to Keynes, he maintained that the sole levy that would enable the Germans to pay reparations to France would be a rise in taxes. Of course, in the flexible exchange rate regime that was then prevailing, the D-Mark would depreciate and the wage rates of German workers expressed in foreign currency would diminish; but the price of German products would diminish in

proportion, so that real wages remained unchanged. It was, however, his analysis of unemployment that made him famous. Following the First World War, unemployment rose in Great Britain and changed in nature: instead of being cyclical, it became permanent. Drawing upon the relation he had put forward between unemployment and the real wage rate, Rueff suggested that this development followed from the emergence of a system of unemployment relief which checked the fall in the money wages despite the existence of an excess labour supply.

The establishment of a more direct link between theory and empirical research involved the development of statistics. Lucien March was the first Frenchman to make Karl Pearson's work known, and he took (*Les principes de la méthode statistique,* 1930) from Pearson three fundamental techniques: the method of moments, the system of curves, and correlation analysis. Marcel Lenoir's 1913 doctoral dissertation *(Etudes sur la formation et le mouvement des prix),* which dealt with price formation and price movements, marked the beginning of econometrics. He not only made careful use of correlation and regression, but he posed, and resolved, the problem of identification. If one had a time series of quantities exchanged and their prices it was possible to plot a path on a graph, but not to interpret this graph as a supply or a demand curve. Lenoir, using moving averages, plotted the long-run trend of cyclical fluctuations. He then calculated regression coefficients and interpreted his results by introducing the idea that short-run variations in prices reflected shifts of the demand curve, while long-term variations were more indicative of shifts in the supply curve and the influence of monetary factors.

Apart from the engineers, French mathematicians took hardly any interest in political economy. Two of them however, Louis Bachelier and Émile Borel, did, at the beginning of the 20th century, make fundamental contributions to the development of economic science. The arguments advanced in Bachelier's *Théorie de la speculation* (1900) lie at the origins of the mathematical analysis of finance: here can be found the essentials of the theory of efficient markets and the premises of the notion of Brownian motion which he developed in 1913. Borel's point of departure is the analysis made by Joseph Bertrand of the game of baccarat in his *Calcul des probabilités* (1889). Bertrand highlighted the existence of a strategic interdependence between the players similar to that which, he suggested, Cournot had wrongly ignored in his analysis of duopoly. But Borel in turn accused Bertrand of overlooking the case where players determined their strategy by drawing lots. He argued that, if one were to reveal the psychological mechanism governing choices, then it had to be connected to the notion of probability: at each moment, each player chooses his or her strategy with a given probability. The player's mathematical hope of gain depends on the way in which the probabilities are allocated to each alternative. In a symmetric game no information can provide one of the players with the certainty of the gain advantage. The best strategy is to distribute probabilities so that one does not lose whatever the opponent does. Borel demonstrated in *La théorie du jeu et les équations intégrales à noyau symétrique* (1921) that a solution exists for a game in which two players could choose between three ways of playing. Nonetheless, it was von Neumann who in 1928 demonstrated at a general level the theorem of the minimax. Jean Ville suggested in 1938 a more simple demonstration, and showed that the result applied to continuous variables.

From 1943 to the Present Day

The publication in the early 1940s of books by Robert Marjolin (*Prix, monnaie et production,* 1941), Maurice Allais (À *la recherche d'une discipline économique,* 1943), François Perroux (La *valeur,* 1943) and by Jacques Rueff (*L'ordre social,* 1945) all testify to a shift in the analyses of French economists. But if they were all certain of the need for a break with traditional liberalism, their work led in different, even contradictory, directions.

Liberals, Keynesians and Institutionalists

If, despite the efforts of Daniel Villey and Louis Baudin, the heritage of French classical liberalism

was fading, after 1940 liberalism experienced a renaissance, but it was a liberalism quite different from that of Molinari and Leroy-Beaulieu. Its most typical representatives, Rueff and Rist, admired Walras for the manner in which he showed that variations in prices always led to equilibrium, since they continued up to the point where they stabilized. René Courtin took up exactly this point in his *Cours de théorie économique* (1950) when he accused Keynes of having assumed absolute rigidity of prices, and of nominal wages in particular. If such a rigidity exists (a doubtful interpretation of Keynes's book), it is never absolute, for while it is capable of explaining unemployment in the short run, it cannot explain its persistence. According to Rueff, the modern social order rests on two institutions: property rights which prevent appropriation by violence, and the market, with its characteristic flexibility of prices which mutually adjust to the point where equilibrium is reached. A property right should be understood as a pool of value, of known volume, which can be filled with whatever wealth offered on the market at the behest of its owner. In so far as the value of this pool corresponds to the value of the goods that it contains, one can say that the right is a real one. But if this is not so, then the right is false. Rights of this sort can be introduced in a number of ways. The simplest example is that of a budget deficit financed by the creation of money. The state, by buying goods or leasing services, creates rights for its creditors. When these expenditures are covered by taxes the rights are real; but if they are not so covered then they are false rights – state creditors hold paper claims to wealth which does not exist. Inevitably, policies of this kind lead to inflation. And in so conducting itself the government weakens the judicial system that protects the social order. Some individuals are not able to provide the rights which they hold with the volume of their choice. The unconditional character of the law is irremediably compromised.

Soon after the publication of the *General Theory,* several works inspired by Keynes appeared, in particular the works of Marjolin (*Prix, monnaie et production,* 1941), Claude Gruson *(Esquisse d'une théorie générale de l'équilibre économique,* 1949) and Alain Barrère (*Théorie économique et impulsion keynésienne,* 1952). They touched on Keynes's work in a very specific manner. Their common problem was the construction of dynamic analysis. They had doubts about the analysis that Keynes had developed in the *General Theory*, but his book had the merit of addressing – even if not fully consciously – the economic problems of growth, and the most fundamental economic policy issue, that of growth coordinated by deliberate and conscious policy. They showed little interest in the models that Modigliani and Hicks had introduced to analyse short-term monetary and budgetary policy. The IS–LM model was for many years neither taught nor discussed in France.

The majority of university economists remained distanced from both liberal arguments and Keynesian ideas. They argued that it was barely possible to understand economic choices without studying its social, cultural and institutional determinants. They argued for a concrete and positive economics closely linked to other social sciences such as sociology and history. The will to renew the link to positive economics was expressed with the foundation in 1950 of the *Revue Économique,* which quickly became the most important of French academic journals. Aftalion was among the founders, alongside historians such as Braudel and Labrousse. This conception of economic science led them to place the study of structure, defined as an ensemble of relations characteristic of a social and economic system – following the example of André Marchal's *Systèmes et structures* (1959) – at the centre of their studies. This method was applied in particular to the analysis of distribution (as in Jean Marchal and Jacques Lecaillon, *La répartition du revenu national*, 1958–70), production structures, spatial organization and the relationships between national economies.

François Perroux played an important role after the Second World War. He created and directed the Institut de Sciences Économiques Appliquées, which for many years was the leading centre for economic research in France. He became a professor at the Collège de France, the most prestigious French scientific institution. Perroux was

open to different influences, and which sometimes appeared to conflict. His first works, in particular his book *La valeur,* revealed the influence Austrian marginalists had played in his thinking. *Économie appliqué*, the journal that he edited, was one of the important channels for the diffusion of Keynes' thinking in France. But his masters were Chamberlin and Schumpeter. He admired Schumpeter as the theorist of innovation, and of creative destruction. What interested him about Chamberlin was the detailed criticism of hypotheses regarding pure and perfect competition. He proposed a general theory of the impact of domination at the level of enterprise, industry and national economy. He saw in this analysis a first and indispensable step towards a much larger synthesis between a theory of the economy and a theory of force, power and of constraints.

And so following the Second World War French economists sought to reconnect with the tradition of positive economics founded with Aftalion and Simiand. This institutionalist project collapsed at the end of the 1960s when the new generation turned to either Marxism or the theory of general equilibrium. Nonetheless, institutionalism has remained an active force within French political economy up to the present day with the Convention School (André Orléan, *Analyse économique des conventions,* 1994) and the theory of regulation (Robert Boyer, *La théorie de la régulation: une analyse critique,* 1986; Boyer and Saillard, *Théorie de la régulation: l'état des savoirs,* 1995). In both schools there is agreement that political economy has to collaborate with other social sciences, history and sociology. The conventionalists are interested in situations where existing prices are insufficient to coordinate the activity of agents on account of uncertainty concerning the future and the quality of products. It is necessary to take account of conventions, understood as legitimate routines of interpretation on the part of agents. The theory of regulation has much larger ambitions: the development of an economic theory which presents an alternative to orthodox theory. Its key concept is the mode of regulation, that is, the manner in which several institutions (the financial system, the wage relation, forms of competition) join together to form a system. Hence the Fordist mode of regulation is characterized by oligopolistic competition, the development of credit, the growth of productivity in mass production and the indexation of wages to gains in productivity. The theory of regulation addresses itself to the description and explanation of different forms of regulation and the specificity of the crises which characterize it.

Reformulations of General Equilibrium Theory

Divisia and Roy had not profoundly modified the basic framework of Walrasian analysis. In 1943 Allais had put forward some new directions for research by introducing intertemporal economies, where each good is defined by the location and date at which it becomes available, and in which there exist markets for all future goods. He demonstrated, making use of Walrasian *tâtonnement,* that the equilibrium was stable. He established the two propositions fundamental to the theory of welfare. In 1947, in *Économie et Intérêt,* he developed a synthesis combining the theory of interest, prices and money. He put forward the first proof of the golden rule. He noted that the existence of transaction costs explained why agents hold money rather than stocks and shares. On this basis he showed that the demand for money is a function of income and of the rate of interest. To illustrate the influence of basic elements of the theory of interest, he introduced a model of overlapping generations. The third fundamental contribution by Allais was the development of a theory of decisions in a state of uncertainty. He showed in *Le comportement de l'Homme rationnel devant le risque* (1953) that, if one wants to account for the behaviour of agents, it is necessary to take account of characteristics of the index of utility other than its average. Finally, in his *La théorie générale des surplus* (1981), Allais put forward a complete modification of the frame of reference: in place of the Walrasian market model he put forward a model of markets founded upon the decentralized search for realizable surpluses.

Debreu was trained as a mathematician; he had been the pupil of Henri Cartan and through him had come under the influence of the Bourbaki group which had an axiomatic approach to

mathematics. It was through the study of Allais's book *À la recherche d'une discipline économique* that he was initiated into the theory of general equilibrium. If Debreu found in his reading of Allais the point of departure for his own studies, the reorientation is significant. Up to that point economic analysis consisted in maximizing differentiable functions and deriving the characteristics of maxima from first-order conditions. Debreu abandoned this approach; differential calculus gave way to topological arguments which quite clearly increased the generality and simplicity of theory. But it was not only the mathematical tools that changed. Allais had maintained that 'in the last analysis it was experience, and only experience, which could determine whether a theory had merit or whether it must be rejected' (Allais 1943, p. 116). In the work of Debreu, the concern for rigour dominates: he stipulated that the axiomatic form of analysis or of theory was, strictly speaking, logically entirely disconnected from its interpretations. In his *Théorie de la valeur* (1954) Debreu took up the analytical framework employed by Allais in 1943. He demonstrated the existence of an equilibrium and established the two theorems of welfare through the use of convex sets. But he refrained from discussing the problem of stability which was central to Allais's preoccupations. The uniqueness of equilibrium posed a problem. At the end of the 1960s it became evident that the hypotheses under which the uniqueness of equilibrium could be established were too restrictive and that it was necessary to make do with an analysis of local equilibrium. Debreu (1970) demonstrated that, using the hypothesis of differentiability, the number of economies that did not have a local equilibrium was 'negligible', that is, 'contained in a closed set of Lebesgue measure zero'. This result, gained by using the concepts and techniques of differential topology, was the origin of the theory of regular economies that Yves Balasko in particular developed.

Following the Second World War the problems of reconstruction, of developing a system of indicative planning, and the management of public enterprises lent Allais, Pierre Massé and their pupils occasion to apply the theoretical propositions that they had elaborated. Among the contributions that French economists made during this period to the theory of the efficient allocation of resources and to the study of public policy, Jacques Drèze (1964) underlined the importance of two themes: the management of public enterprises and the analysis of the conditions under which the accumulation of capital is socially effective.

Edmond Malinvaud (1953) explicitly introduced time into the model of general equilibrium. From this he derived an analysis of the determination of the rate of interest and the meaning that it gives to the proposition that the rate of interest is equal to the marginal productivity of capital. One can only regret that the economists who became involved in the controversy that led to the theory of capital did not always record the results that they arrived at.

Marcel Boiteux (1956) suggested a new approach to the management of public monopolies constrained by budgetary equilibrium. He sought to define a rule for the management of public monopolies by adding to natural connections a new constraint: the budgetary equilibrium. He then defined the shadow prices which were the solution to the problem. Public monopolies should maximize their profits in terms of these shadow prices. The gap between real prices and shadow prices is proportional to the inverse of the price elasticity of compensated demand. While Dupuit and Colson referred to marginal costs, Boiteux took account of shadow marginal costs and prices.

What remains to be determined is whether the enterprise or the regulator is the better at determining tariffs. Jean Tirole and Jean-Jacques Laffont analysed systematically this type of problem by using the theory of contracts. The central idea is that information at the disposal of the managers of a public monopoly is greater than that available to the regulator. It is therefore necessary to determine the nature of the contract which the regulator is able to propose to the enterprise to minimize the costs of production of the good which it produces, while explicitly taking account of the capacity of the agent to manipulate the information.

In Debreu's model, all agents have, *ab initio,* access to a complete system of forward markets and adjustments are made solely by price.

All contracts are concluded on the starting date; there is no incentive to reopen markets at a later date. The model is essentially atemporal; the role of money cannot be explained, nor the existence of a market for stocks nor the underemployment of resources. Lindahl and Hicks suggested that a partial equilibrium framework was appropriate for dealing with this kind of problem. Michel Grandmont, in a series of articles published in the course of the 1970s, took up and then systematically developed this notion by assuming that agents formed, at every moment, expectations of the future states of the economy that were not necessarily realized. It was in this framework that, in the 1970s, Jean-Pascal Benassy, Drèze, Malinvaud and Yves Younès built their theory of disequilibrium. More recently, this framework was used to study the relations between value and money (Grandmont, *Money and value,* 1983), between competition and underemployment (Claude D'Aspremont, Louis Gérard-Varet, Rodolphe Dos Santos, *On Monopolistic Competition and Involuntary Unemployment,* 1990, and Benassy, *The Economics of Imperfect Competition and Underemployment,* 2002) and rational expectations (Roger Guesnerie, *Assessing Rational Expectations*, 2001).

Until the 1970s, French economics had a flavour of its own with engineer–economists interested in planning and the management of public enterprises, and with many professors still following the French institutionalist tradition. Thereafter, this distinctiveness disappeared and, with the exception of the Regulation School, French economists became thoroughly integrated into an international economics profession.

See Also

- ▶ Allais, Maurice (born 1911)
- ▶ Allais Paradox
- ▶ Aupetit, Albert (1876–1943)
- ▶ Bachelier, Louis (1870–1946)
- ▶ Bertrand, Joseph Louis François (1822–1900)
- ▶ Braudel, Fernand (1902–1985)
- ▶ Debreu, Gerard (1921–2004)
- ▶ Divisia, François Jean Marie (1889–1964)
- ▶ Divisia Index
- ▶ Gérard-Varet, Louis-André (1944–2001)
- ▶ Gibrat, Robert Pierre Louis (1904–1980)
- ▶ Juglar, Clément (1819–1905)
- ▶ Laffont, Jean-Jacques (1947–2004)
- ▶ Leroy-Beaulieu, Pierre-Paul (1843–1916)
- ▶ Massé, Pierre (1898–1987)
- ▶ Perroux, François (1903–1987)
- ▶ Roy, René François Joseph (1894–1977)
- ▶ Sauvy, Alfred (1898–1990)
- ▶ Walras, Léon (1834–1910)

Bibliography

Allais, M. 1943. *À la recherche d'une discipline économique. Première partie, l'économie pure.* Saint-Cloud, chez l'Auteur, deuxième édition sous le titre *Traité d'économie pure.* Paris, Imprimerie Nationale, 1952; troisième édition, Paris: Clément Juglar, 1994.

Aréna, R. 2000. Les économistes français en 1950. *Revue économique* 51: 969–1007.

Boiteux, M. 1956. Sur la gestion des monopoles publics astreins à l'équilibre budgétaire. *Econometrica* 24: 22–40.

Debreu, G. 1954. *Theory of value*. New York: Wiley.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38: 387–392.

Divisia, F. 1951. *Exposés d'économique, Introduction générale. L'apport des ingénieurs français aux sciences économiques*. Paris: Dunod.

Dockès, P., L. Frobert, G. Klotz, J.-P. Potier, and A. Tiran. 2000. *Les traditions économiques françaises*. Paris: CNRS éditions.

Drèze, J. 1964. Some postwar contributions of French economists to theory and pubic policy: With special emphasis on problems of resource allocation. *American Economic Review* 54(4), Part 2: Supplement: Survey of foreign postwar developments in economic thought, 1–64.

Faccarello, G. (ed.). 1998. *Studies in the history of French political economy, from Bodin to Walras*. London/New York: Routledge.

Greffe, X., J. Lallement, and M. De Vroey (eds.). 2002. *Dictionnaire des grandes œuvres économiques*. Paris: Dalloz.

Le Van-Lemesle, L. 2004. *Le Juste ou le Riche. L'enseignement de l'économie politique, 1815–1950*. Paris: Comité pour l'histoire économique et financière de la France.

Malinvaud, E. 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21: 233–268.

Steiner, P. 2000. La Revue Économique, 1950–1980: la marche vers l'orthodoxie économique. *Revue Économique* 51: 1009–1058.

Zylberberg, A. 1990. *L'économie mathématique en France, 1870–1914*. Paris: Economica.

France, Economics in (Before 1870)

Alain Béraud and Philippe Steiner

Abstract

From the late 17th century onwards, French economists were major contributors to the rise of economic liberalism, developing many of the analytical tools of political economy. After the Revolution, their major concern was the growth and stability of what they called 'industrial society'; and a distinction arose between those who claimed that such a society needed to be regulated (the Saint-Simonians) and those in favour of a more decentralized and market-oriented system. After 1848, French economists became deeply involved in the struggle against socialism, and devoted a great deal of energy to the diffusion of sound principles of political economy.

Keywords

Animal economy; Arrow, K. J; Bastiat, F; Batbie, A; Bicquilley, C. -F; Blanc, L; Boisguilbert, P; Calculus; Cameralism; Canard, N; Cantillon, R; Chevalier, M; Child, J; Cobden–Chevalier Treaty; Colbert, J. -B; Condillac, E. B; Condorcet, M. de; d'Alembert, C; Decreasing returns to capital in agriculture; Diderot, D; Division of labour; Dupuit, A. -J; Economic governance; Entrepreneurship; Fiduciary money; Forbonnais, F. V. de; France, economics in; Garnier, J; Gournay, V. de; Impossibility theorem; Industrialism; Isnard, A. -N; Law, J; Le Mercier de la Rivière, P. -P; Mably, Abbé de; Mercantilism; Mirabeau, Comte de; Molinari, G. de; Montchrestien, A. de; Montesquieu, Comte de; Mutualism; Natural price; Necker, J; Net product; Overproduction; Physiocracy; Political economy; Probability; Proportionate price; Protectionism; Proudhon, P. -J; Public debt; Quesnay, F; Ricardo, D; Right of association; Roederer, P. -L; Rossi, P; Rousseau, J. -J; Saint-Simonians; Say, J. -B; Single tax; Sismondi, J. C; Social mathematics; Socialism; Subjective theory of value; Taxation in kind; Turgot, A. R. J; Vauban, S

JEL Classifications

B1

From the End of the 17th Century to 1755

The term *économie politique* first appeared in French in Antoine de Montchrestien's *Traicté de l'œconomie politique* of 1615. However, during the 17th century there was no French counterpart to English mercantilist thought, nor the kind of economic administration formed on cameralist principles found in Austria and Germany, despite Colbert's attempt to promote the wealth and power of the monarchy through the regulation of commerce. Censorship and the weakness of the French merchants as a class could explain this situation. By the end of the 17th century reflection on economic matters was just beginning, and the monarchy was increasingly conscious of the gravity of the problems that recurrent dearth and high levels of debt represented. This created the conditions for the questioning of economic policy with respect to both the provisioning of markets and taxation.

Vauban argued in his *Dixme royale* (1707) that the principal cause of the monarchy's economic distress was the way its fiscal system was organized. Taxation, he wrote, should be raised in kind as a proportion of the gross yield from the annual harvest. Such a tax would therefore be proportional to agricultural wealth. For commerce and industry he anticipated light taxes that could be passed on in trade.

Boisguilbert's proposal (*Le détail de la France,* 1695) goes much further, even though his attention was likewise directed to taxation. His theory of markets derived from Jansenist moral philosophy, according to which a society in which behaviour was founded upon interests would also be ordered in the same way as a society

composed of charitable and pious people. Boisguilbert endorsed laissez-faire as the sole condition permitting the emergence of the 'proportionate price', a price at which each gained from participating in exchange and in which each party to the exchange adhered to his budget constraint. He argued that both good and bad harvests disrupted economic activity because they would bring about violent price changes if, as was then the case in France, free competition were absent. Since the wheat market determined the level of agents' revenues (the remuneration of agricultural capital, the payment of rents), variations in the price of wheat affected other markets. Moreover, the price of wheat was vital to the subsistence of populations. Expectations on the part of agents, whether justified or not, disturbed the economy, and government intervention was not capable of stabilizing the market since such intervention was in turn perceived to be the sign of an even more serious crisis.

After the death of Louis XIV in 1715, the regent accepted John Law's arguments concerning financial policy. According to Law (*Considérations sur le commerce et l'argent,* 1720), France's poor economic performance was due to an inadequate money supply. In 1716, he founded a bank which had the creation of paper money as its principal function; this paper money was supposed to substitute for coins and to permit a refinancing of government debt. Here Law's ideas were at variance with those of Boisguilbert, but Law also went on to argue that money could also be backed by land or by shares, that is, by productive capital. These ideas were given shape with the formation of a commercial company that was granted an exclusive right to trade with Louisiana. The company's shares could be purchased only with *billets d'Etat* (government securities) at their face value instead of being discounted about 70 per cent, but the public could hope for capital gains if the company's trade was well managed. The company gained in this way an exclusive right to the exploitation of vast wealth, and the state transformed its floating debt into long-term debt with a lower interest rate. The merging of the company and the bank permitted monetary expansion and at the same time boosted the value of the company's own shares. At the end of 1719 Law became Comptroller General of Finance – money issue was strong (around a million *livres)* and the rate of interest touched a low point of two per cent. The price of the company's own shares was stabilized by an office which intervened in the market. The system collapsed as soon as agents sought to exchange their shares and securities for cash. Law's collapse had a lasting impact. The chance of modernizing the public finances had been missed, and for the entire 18th century the collapse weighed heavily on the capacity of the French monarchy to finance its military conflicts with Britain. In addition, a marked suspicion of fiduciary money and banking prevailed right up to the Revolution.

Discussion of monetary matters and Law's system continued in the early part of the 18th century, but gave way to an interest in commerce from the perspective of the legislator, as for instance in Richard Cantillon's *Essai sur la nature du commerce en général* (written around 1728–30 and published in 1755) and Jean-François Melon's *Essai politique sur le commerce* (1736). Cantillon's text is the more notable of the two on account of his theory of price (measured in land) and his general theory of the circulation of goods founded upon the behaviour of the entrepreneur. The theory of the balance of trade is modified by taking account of the value in land of the products exchanged, and Cantillon associates with it an automatic equilibrating mechanism mediated by modifications to the expenditures of landed proprietors. The science of *commerce politique* was given a decisive boost in 1751 with Vincent de Gournay's accession to the post of Supervisor of Commerce. The intention was that France should follow the example of England in supporting mercantile activity, but Gournay's economic thinking was not that original: it remained close to the brand of mercantilism advanced by Josiah Child and which saw in a low rate of interest the best way of promoting commerce. His significance, rather, lay in the fact that he gathered around himself young administrators (such as Véron de Forbonnais, the abbé Morellet, and Turgot) who would be influential up to the time of the Revolution.

The science of commerce that crystallized in de Gournay's writings and those of his group, or in Montesquieu's *Spirit of Laws,* can be characterized by four features. First, trade is composed of flows of goods between nations which exchange their surplus thanks to the practical knowledge of traders. Second, trade depends upon self-interested behaviour, and it implies that the trader has an interest, both economic and symbolic, in keeping to his particular station in life rather than in achieving nobility. Third, trade is the most important form of economic activity. And fourth, the particular interest of the trader could be opposed to that of the state.

1756–1789: From Physiocratic *Philosophie économique* to Condorcet's Social Mathematics

From 1750, economic publications multiplied and this growth accelerated in the years leading up to the Revolution. New contributors to the genre emerged with François Quesnay and the Physiocrats during a troubled political period including the Seven Years' War (1756–63) and the Treaty of Paris, under which a large part of the French colonial empire was lost.

Diderot and d'Alembert's *Encyclopédie* presented Forbonnais with the opportunity of writing a series of entries which were then brought together in his influential *Eléments du commerce* (1754). This publicized the views of the group around de Gournay on the importance of monetary flows and a low rate of interest. But there were two other important contributors to the *Encyclopédie.* Rousseau argued that the General Will was the first principle of political economy and the basic rule of government. This proposition opposed republican virtue to wealth and interested behaviour. The abbot Mably took this argument up in criticism of the Physiocrats (*Doutes présentés sur l'ordre légal et essentiel des sociétés politiques,* 1767). The same argument was revived during the Revolution, when the most radical of the Montagnards reclaimed for themselves ancient republican egalitarianism in order to promote the right of property and economic development through the market.

Quesnay came to political economy from medicine. There he had encountered the then contemporary notion of animal economy – economy understood as a harmonious organization of diverse phenomena which came together in one coherent whole: the body. He transferred this notion, as was fashionable at the time, to the level of the state so that he was able to talk of economic government, a concept vital to the presentation of his ideas. The task of economic government was to administer resources – men, land, money – in such a way that the nation would enjoy abundance; under-employment of resources was not to be attributed to individuals, but to the errors of economic government. According to Quesnay (*Grains,* 1757), economic government should leave the decision of what is best in matters of culture or trade to the interested behaviour of men. It should limit itself to providing an institutional context favourable to interested behaviour; commercial freedom and a predictable tax levied upon the net product (and not on the gross product as in Vauban) so that productive capital might be maintained. The latter was later elevated to the status of the central variable in the economy since the amount of the net product is always fixed as a proportion of farmers' circulating capital.

Quesnay elaborated the advantages of free trade in the market for wheat in arguments that Dupont de Nemours and Turgot then adopted. He explained how free trade blunted brutal market fluctuations – a phenomenon noted by Gregory King and elaborated by Charles Davenant (*Essay upon the Probable Methods of Making a People Gainers in the Balance of Trade,* 1699) in the 17th century – by allowing compensating adjustments between nations. The consumer enjoyed the benefits of more stable prices. The producer who would benefit from a better price will be prompted to produce more – so long as the price did not fall too far as a consequence of a good harvest. These interests conjoin those of the consumer (in the security of provision) and those of the state (enhanced wealth and increased fiscal returns).

In 1758 Quesnay converted to his camp Count Mirabeau, whose *L'ami des hommes* (1758) on population and commerce had been well received. Their subsequent close collaboration led to the major doctrinal publications of Physiocracy – *Théorie de l'impôt* (1760) and *La philosophie rurale* (1763) – in which Quesnay elaborated his idea of the single tax payable by sole proprietors on the grounds that they were the sole recipients of agricultural rent. But the theoretical landmark of this period is the *Tableau Economique*, which appeared in different versions between 1758 and 1767. There are echoes in the *Tableau* of Cantillon's approach, his text having circulated in manuscript before 1755. Flows between rural and urban classes are conceived at the highest level of abstraction so that the relation of these classes to each other might be clearly demonstrated. The key difference is that Cantillon was interested in monetary phenomena and commercial uncertainty, matters neglected by Quesnay.

In the initial versions of the *Tableau*, Quesnay showed how landowners' expenditures made possible the circulation of the wealth produced by farmers and artisans. The later versions, more 'macroeconomic' in form, showed under what conditions the monetary expenditure of a society restricted to three classes (farmers, landowners and artisans) allowed the reproduction of the conditions of agricultural wealth at an optimal level. This final version of the *Tableau* also allowed the impact on the amount of the net product of accrued luxury expenditures, or of indirect taxation, to be studied; it hence made possible an estimation of their importance to the nation as a whole.

The Physiocratic School gained in importance during the 1760s and played a role in economic administration. In 1764–5 the Comptroller General, Bertin, liberalized trade in wheat and in flour; together with Turgot, Inspector in Limousin, and Pierre Paul Le Mercier de la Rivière, Inspector in the Antilles, the highest reaches of administration opened up to Physiocracy. The doctrine spread abroad: to Baden, Austria, Poland, Russia and Sweden. However, a succession of poor harvests in the later 1760s put an end to those tentative efforts at trade liberalization. Quesnay lost interest in political economy; the baton was taken up by a small number of writers, among whom Turgot was pre-eminent.

Turgot is close to Physiocracy, but he differs in theoretical points and practical matters. He was close in so far as he was a strong advocate of a complete freedom of trade, distancing himself from Gournay's slogan 'liberty and protection'; and he adopted Quesnay's analysis of the price of wheat in respect of the theory of the net product and the single tax. But Turgot never made use of the *Tableau Economique;* he was, he said, happy to employ its metaphysics, meaning the competitive process upon which it was founded.

Turgot's originality is evident from his *Réflexions sur la formation et la distribution des richesses,* published in 1766 in the Physiocratic journal *Ephémérides du citoyen,* and can also be appreciated from many of his writings of this period that were either never completed, or remained unpublished, such as his essay *Valeur et monnaie*. His approach is based upon sensualist philosophy, and this orients him to a subjective theory of value and utility. The economic thought of abbé de Condillac, the principal theorist of sensualism in France, was similar in this respect, for in his *Le gouvernement et le commerce considérés relativement l'un à l'autre* (1776) Condillac defined value in terms of judgement and opinion made in respect of the scarcity and utility of a good – combining this with a more thorough study of the competitive process.

This led Turgot to a number of significant findings: the formation of markets upon the foundation of mutual interest between buyers and sellers constrained by transport costs (*Foires et marchés,* 1757); a theory of price (estimated value) proceeding from a discussion of the scarcity (quoted value) of a good for parties to an exchange – although Turgot stopped at two agents and two goods (*Valeur et monnaie,* 1769); the justification for interest upon loans and its determination according to market forces (*Mémoire sur le prêt à intérêt,* 1770); a theory of the formation of a uniform rate of profit, or a stable hierarchy of such rates (*Réflexions*, 1766). If one adds to this list the discovery of the principle of decreasing returns to capital in agriculture it is clear that

Turgot's theoretical contribution was a considerable one, especially in view of the fact that he had heavy responsibilities in his various administrative posts – as Inspector in Limoges (1761–74), then Navy Minister (1774), and finally Comptroller General for Louis XVI (1774–6).

In this last appointment, together with a small number of loyal supporters (Dupont de Nemours, Condorcet) Turgot worked to re-establish the freedom of trade in grain, which gave rise to a dispute with Jacques Necker (*Sur la législation et le commerce des grains,* 1775), Necker opposing to Turgot's liberalism a more flexible and pragmatic conception of the administration of trade for which the anticipations and beliefs of agents were vital, a factor neglected by Turgot.

Political economy thus assumed an explicitly political dimension. For Quesnay and Le Mercier de la Rivière (L'*ordre naturel et essentiel des sociétés politiques,* 1767) the community of economic interests shared by different groups secured the harmony of the social body, provided that the legislator surrounded himself with experts in the science of economics. Mirabeau and Turgot considered that landed proprietors represented the general interest and should determine the level of taxation in local assemblies. This connection between property, taxation and the citizenry would play an essential role in the course of the Revolution. This connection is also the basis upon which a general science of the social was conceived (the moral or political sciences according to the abbé Baudeau, and social science as in Sièyes, Condorcet or Roederer) in which political economy took its place alongside ethics, politics and jurisprudence. It was in this form that political economy was first institutionalized in the classes on moral and political sciences at the Institut (1795).

We should also take note of a specific development owed to the presence of Condorcet, a mathematician of the first rank, in Turgot's entourage. Condorcet's interest in public affairs during the Revolution gave rise to his essays on social mathematics which inserted calculus and the theory of probability into social science with respect to issues such as insurance or the rate of interest on loans. Quite remarkable is the result obtained by Condorcet in respect of the determination of truth on the part of a jury or assembly when there are several votes and more than two choices. Condorcet formulated the result which Kenneth Arrow demonstrated in 1951 as the 'impossibility theorem'. But for the time being, this avenue remained undeveloped, apart from the work of isolated scholars like Achylle-Nicolas Isnard (*Traité des richesses,* 1781), Nicolas Canard (*Principes d'économie politique,* 1801) or Charles-François Bicquilley (*Théorie élémentaire du commerce,* 1804). Say rejected it quite explicitly.

F

1800–30: Say, the Saint-Simonians and the Industrial Order

Physiocracy continued to play a role during the revolutionary period. A number of followers had been shaped by this doctrine, and this remained true even of those who had distanced themselves on central points, such as the abbé Sièyes, Roederer or Condorcet. However, the diffusion of the *Wealth of Nations* profoundly altered the way in which the economy was conceived in France. Two authors symbolize this progression: Jean-Baptiste Say (*Traité d'économie politique,* 1803) and Jean-Charles Simonde de Sismondi (*De la richesse commerciale,* 1803). Despite their evident indebtedness to Quesnay and Turgot, many traces from these authors remaining in their writings, they founded their political economy upon the *Wealth of Nations*, Germain Garnier's influential translation being published in 1802. For Say and Sismondi, Smith had highlighted two salient points. The first was that the industrial producer acquired his social independence thanks to the market. He no longer depended upon a person of influence (such as a rich landed proprietor) but on a collection of purchasers. The second was that the level of economic activity did not depend on expenditures, but on the quantity of capital. In this respect the social and political dimension of political economy came to the fore in a conception of a new type of society which Say called 'industrial society'.

Say's political economy is characterized by the manner in which he orders his material by the

tripartite schema of production, distribution and consumption. He did more than simply put Smith's ideas in order; he modified both Smith's ideas and those of his British interpreters. Say followed the tradition of Turgot and Condillac. His theory of value is based on utility, not labour. He thus rejected the opposition of natural to market price, in the last editions of the *Traité* considering only market price. Say's theory of production minimizes the role of the division of labour. He argues that the progress of wealth arises from the introduction of new machines incorporating scientific knowledge which places at the disposal of producers the free forces of nature, thereby reducing costs of production. The theory of distribution is entirely based on relations between supply and demand among different categories of the suppliers of productive services, including those of entrepreneurs.

Say's name is firmly linked to two fundamental contributions: his formulation of the law of markets and his analysis of the role of the entrepreneur. The latter played a significant part in his theory. The entrepreneur coordinates the employment of productive services within an enterprise and links different markets (for final goods and for productive services). In this respect Say's entrepreneur, as in Cantillon, is the economic agent who confronts the uncertainties involved in market transactions.

Say argues that value depends upon utility and is the measure of wealth. From 1815 Say encountered criticism on these two points from Ricardo, and never managed satisfactorily to meet the criticism that the fall in value of a good consequent upon technical progress cannot at the same time indicate that the society is richer (a given amount of utility being obtained at a lower cost) and also poorer (since value has diminished). In this debate Say had trouble in defining a theoretically founded position which was not a reformulation of the Ricardian theory, including here the theory of rent. The difference in method is certainly here more marked and on this point Say received support from Sismondi (*Nouveaux principes d'économie politique,* 2nd edition, 1826). But they were not in agreement on the implications of the law of market opportunities and on the interpretation of the English industrial crisis of 1825: for Say, it resulted from excessive credit being extended by banks, while Sismondi saw it as a crisis of overproduction originating in the growth of production exceeding that of consumption.

In France the debate on value took a distinctive course. Rossi, the successor to Say at the Collège de France, rapidly abandoned the position of his predecessor and moved nearer that of Ricardo. He also elaborated a methodological synthesis which distinguished between a pure and abstract economics in the fashion of Ricardo and an applied political economy influenced by institutional and political context. Most importantly, however, following on from Rossi, Dupuit criticized Say's position: the value of a good was not measured by its utility, instead one might measure utility by the maximum sacrifice a purchaser was prepared to make to obtain it.

Beyond these theoretical debates, the political economy of Say and his successors bore upon the nature of society. The doctrine of industrialism expressed the idea that modern society depended upon the mastery of man over nature thanks to science and technology on the one hand, and social science on the other. Industrialism endorsed and promoted industry, the social independence produced by the market and the reconfiguration of the political sphere, where the state played a diminished role, permitting agents to decide what was best for themselves while it also assigned a greater role to industrial classes in the representation of the citizenry. During the 1820s this doctrine divided into two paths: the liberal industrialism of Say, Charles Dunoyer and Charles Comte separated from the organized industrialism of Henri-Saint-Simon, Auguste Comte and the Saint-Simonians. This latter tendency argued that the market was not an institution adequate to the effective redistribution of resources, as periodic economic crises showed. It was the same with the hereditary transmission of property; in its place, industrialism envisaged a centralized and rational organization of economic activity. In addition, it asserted that industrial society could not be based simply on selfish interest and the doctrine of utility, but had need of a

moral or religious dimension. Here we are already approaching socialist theses that flourished during the 1840s.

This opposition assumed particular force with the link that developed between organized industrialism and a new social category, that of the engineer. Since the 18th century France had provided itself with a corps of engineers charged with the provision and maintenance of infrastructure (bridges, roads and canals), mines and defence. These engineers were selected for their abilities in mathematics, and they employed this in a profession placed between technology and economy. 'Engineer economists' (Etner 1987) created a link between political economy and mathematics in economic calculation. This is evident in the work of Dupuit, who calculated the utility of infrastructure, and expounded the principle that a tariff should be charged according to the gain that a user enjoyed. Antoine-Augustin Cournot (*Recherches sur les principes mathématiques de la théorie des richesses,* 1838) was himself a pure mathematician. While he developed an economic approach in respect of theses expounded by Rossi on the value of exchange, he remained, as a writer, isolated in his use of mathematics and also on account of his critique of free trade. His work was hardly read by his contemporaries.

1830–70: The French Classical Liberal School, Socialism and the Teaching of Political Economy

Say dedicated much of his life to teaching political economy: at the Athénée royal (1815–19), at the Conservatoire des arts et métiers (1819–32) and finally at the Collège de France (1830–2). The importance that Say attached to teaching political economy derived mainly from his adherence to Enlightenment philosophy, according to which human misfortune resulted from ignorance of the laws of nature and of society, and from the ascendancy of doctrines which prevented individuals from daring to think for themselves. It also followed from his own economic theory, for he maintained that scientific knowledge was among the productive services that the entrepreneur had to bring together so that he might serve the public effectively.

This perspective came to be of importance in the debate with Ricardo. Say did not neglect theory, and he sought to develop it (the law of markets, the theory of value, the theory of productive services and so on), but he considered that the essentials were already understood. Republican in outlook, Say saw in political economy the means to bring about a more efficient society, one in which there was greater justice because it was more egalitarian. The diffusion of a liberal credo favourable to commercial freedom, free trade and reduced taxation was therefore important. Agreement among economists provided a secure foundation for the production of a body of ideas appropriate for public instruction. Ricardo's theoretical refinements, which he did not himself think had practical consequences, brought about disagreements which alienated readers from political economy and its applications, as shown by the jibes against economists of François Ferrier, a customs official and defender of the balance of trade (*Du gouvernement considéré dans ses rapports avec le commerce, ou de l'administration commerciale opposée aux économistes du 19ème siècle,* 1804 and 1822.)

After the death of Say in 1832 this conception of political economy was epitomized in the various institutions around which liberals organized themselves. In 1832 François Guizot re-established the Académie des sciences morales et politiques that Bonaparte had suppressed; in 1842 economists founded the Society for Political Economy so that they might there discuss theory and policy; the publisher Guillaumin saw that their work was published (the *Collection des principaux économistes* in 1842 and then, in 1852–3, the remarkable *Dictionnaire de l'économie politique).* Finally, liberal economists founded a journal, the *Journal des économistes,* which was published from 1841 right up to the French military collapse in 1940.

The initial aim of the *Journal des économistes* was the diffusion of economic theory, thought to be already complete, so that it might lead to practical ends. The contemporary problem appeared to relate to the forms of association between workers

and capitalists, and support for a spirit of enterprise that had not brought about all its anticipated benefits. Frédéric Bastiat led a powerful campaign on behalf of free trade, seeking to create in France a movement which was the equivalent of Cobden's Anti-Corn Law League. The struggle against socialism was not therefore a priority for liberal economists in dialogue with 'social reformers', notably with Pierre-Joseph Proudhon who, thanks to his relationship with Joseph Garnier, then director of the *Journal*, was regarded a part of the circle of economists and published his *Contradictions économique* with the publisher Guillaumin. It is true that he was similar to them on one count – the defence of freedom – and which he sought to reflect in mutualism, one of the forms of association in question. Matters quickly changed in 1848; the suppression by the new authorities of the chair of political economy at the Collège de France profoundly upset economists, who set to work in support of its re-establishment; and they opposed many projects developed at this time, such as the 'right to work' and national workshops, which generally promoted the centralized regulation of economic activity. Besides writing in support of property and social order, the economists (especially Michel Chevalier and Joseph Garnier) opposed the ideas of Louis Blanc: remunerating work independently of its productive contribution, as in the national workshops, created a problem with incentives. Nevertheless, the *Journal des économistes* saw its principal adversaries as ignorance of the principles of political economy, protectionist prejudices, and socialist illusions. Bastiat developed this idea on his *Sophismes économiques* (1845). Socialism and protectionism were conceived as equivalent, for both involved despoliation, an involuntary transfer of resources which impoverished society to the advantage of one particular section of that society.

The creation of the Empire in 1851 opened up a cleavage among the economists. The most liberal among them, such as Gustave de Molinari, left the country, while others furthered their industrial ideas and political careers, like Chevalier, who became a Privy Councillor and personal Councillor to Napoleon III. From this position he was able to promote the central idea of liberal economics with the signature of the Cobden–Chevalier Treaty on free trade in 1860. During the Empire period there were additional measures that conformed to liberal ideas, such as restoring the right of association to workers in 1864 and furthering education in political economy. Hitherto it had been taught only in several specialized institutions (the Conservatiore, the Collège de France, and the Ponts et Chaussée), but from 1860 public education in political economy began in the provinces, and in Paris in the law faculty with a course given by Anselme Batbie. However, the development of teaching in political economy really began to develop only with the reform of the teaching of law in 1877.

See Also

- ▶ Boisguilbert, Pierre le Pesant, Sieur de (1645–1714)
- ▶ Canard, Nicolas-François (c1750–1833)
- ▶ Cantillon, Richard (1697–1734)
- ▶ Chevalier, Michel (1806–1879)
- ▶ Colbert, Jean-Baptiste (1619–1683)
- ▶ Condillac, Etienne Bonnot de, Abbé de Mureau (1714–1780)
- ▶ Condorcet, Marie Jean Antoine Nicolas Caritat, Marquis de (1743–1794)
- ▶ Dupuit, Arsene-Jules-Emile Juvenal (1804–1866)
- ▶ Forbonnais, François Véron Duverger de (1722–1800)
- ▶ Isnard, Achylle Nicolas (1749–1803)
- ▶ Law, John (1671–1729)
- ▶ Mercier De La Rivière, Pierre-Paul (Mercier or Lemercier) (1720–1793/4)
- ▶ Physiocracy
- ▶ Quesnay, François (1694–1774)
- ▶ Saint-Simon, Claude-Henri (1760–1825)
- ▶ Say, Jean-Baptiste (1767–1832)
- ▶ Sismondi, Jean Charles Leonard Simonde de (1773–1842)
- ▶ Turgot, Anne Robert Jacques, Baron de L'Aulne (1727–1781)

Bibliography

Béraud, A. and Steiner, P., eds. 2004. L'économie politique néo-smithienne sur le Continent: 1800–1848. Special issue of *Œconomia* 34.
Breton, Y., and M. Lutfalla (eds.). 1991. *L'économie politique en France au 19ème siècle*. Paris: Economica.
Charles, L., P. Lefèvre, and C. Théré (eds.). 2007. *Commerce, population et société autour de Vincent de Gournay*. Paris: Ined.
Dockès, P., et al. (eds.). 2000. *Les traditions économiques françaises: 1848–1939*. Paris: Cnrs éditions.
Etner, F. 1987. *Histoire du calcul économique en France*. Paris: Economica.
Faccarello, G. (ed.). 1998. *Studies in the history of French political economy*. London: Routledge.
Faccarello, G. 1999. *The foundations of Laissez-faire: The economics of Pierre de Boisguilbert*. London: Routledge.
Hollander, S. 2005. *Say and the classical Canon in economics: The British connection in French classicism*. London: Routledge.
Kaplan, S. 1976. *Bread, politics and political economy in the Reign of Louis XV*. The Hague: Martinus Nijhoff.
Le Van-Lemesle, L. 2004. *Le juste et le riche: L'enseignement de l'économie politique 1815–1950*. Paris: Comité pour l'histoire économique et financière de la France.
Perrot, J.-C. 1992. *Une histoire intellectuelle de l'économie politique (XVII-XVIII siècle)*. Paris: Ehess.
Potier, J.-P., and A. Tiran (eds.). 2003. *Say: nouveaux regards sur son œuvre*. Paris: Economica.
Steiner, P. 1998a. *La 'Science nouvelle' de l'économie politique*. Paris: Presses universitaires de France.
Steiner, P. 1998b. *Sociologie de la connaissance économique: Essai sur les rationalisations de la connaissance économique (1750–1850)*. Paris: Presses universitaires de France.
Vatin, F. 1998. *Economie politique et économie naturelle chez Cournot*. Paris: Presses universitaires de France.
Whatmore, R. 2000. *Republicanism and the French Revolution: An intellectual history of say's political economy*. Oxford: Oxford University Press.

Franchising

Francine Lafontaine

Abstract

Franchising typically refers to contractual relationships between legally independent firms, where one firm pays the other for the right to operate under the latter's brand, or sell its product, in a given location and time period. Franchised firms account for a large portion of commerce in the United States and around the world. The economics literature on franchising has focused mostly on why and how firms franchise, emphasizing incentive or opportunism issues on the part of franchisees and franchisors to explain various aspects of the relationships. Empirical findings have confirmed the importance of such issues in shaping these contractual relationships.

Keywords

Antitrust; Business-format and traditional franchising; Chain structures; Contract enforcement; Franchising; Industrial organization; Principal and agent; Risk; Royalties; Sharecropping; Vertical integration

JEL Classifications

L22

Franchising typically refers to contractual relationships between legally independent firms under which one of the firms, the franchisee, pays the other firm, the franchisor, for the right to sell the franchisor's product and/or the right to use its trademarks and business format in a given location and for a specified period of time.

According to the *American Heritage Dictionary of the English Language,* the word 'franchise' comes from the old French word *franche,* meaning free or exempt. In medieval times, a franchise was a right or privilege granted by a sovereign power – king, Church, or local government – to engage in activities such as building roads, holding fairs, organizing markets, or to maintain civil order and collect taxes, in a particular location and for a certain period of time. The grantee was typically required to pay a share of its product or profit to the sovereign power for this right or privilege. That payment was called a *royalty*, a term we still use today.

Governments still grant franchises in certain industries, such as the cable television industry

(see for example Zupan 1989; Prager 1990) and highway construction projects (see Engel et al. 2001). The word 'franchise' is used also in the sports industry to refer to the right to operate a team in a particular locale. Most commonly, however, the term refers to the type of ongoing business relationships defined above.

In the United States, the Federal Trade Commission (FTC) has jurisdiction over federal disclosure rules for franchisors. It requires three conditions for a business relationship to be deemed a franchise and thus subject to these rules. First, the franchisor must license a trade name and trademark that the franchisee operates under, or the franchisee must sell products or services identified by this trademark. Second, the franchisor must exert significant control over the operation of the franchisee or provide significant assistance to the franchisee. Third, the franchisee must pay at least 500 dollars to the franchisor at any time before or within the first six months of operation (see Disclosure Requirements and Prohibitions concerning Franchising and Business Opportunity Ventures, CFR, Title 16, Part 436). Authorities outside the United States, including Australia, Canada, and the European Union, typically rely on similar criteria.

Franchise agreements take one of two forms: business-format franchises, where the relationship 'includes not only the product, service, and trademark, but the entire business format itself – a marketing strategy and plan, operating manuals and standards, quality control, and continuing two way communication' (U.S. Department of Commerce 1988, p. 3) and product and trade name or traditional franchising, where franchised dealers 'concentrate on one company's product line and to some extent identify their business with that company' (1988, p. 1). The latter include car dealerships, petrol stations, and bottlers. Several countries, however, exclude these from their franchise statistics.

In 2001, the revenues of franchised chains in the United States were estimated at 1.37 trillion dollars or 13.6 per cent of GDP (Blair and Lafontaine 2005, p. 26). In retailing, it is estimated that about one-third of each dollar of sales is achieved via franchised chains. Three-quarters of these sales occur in traditional franchise outlets. Business-format franchising, however, accounts for the majority of jobs and outlets: of the more than 750,000 franchised establishments in the United States in 2001, 620,000 were associated with the 2,500–3,000 business-format franchisors in the economy. Thus business-format franchising accounted for 4.3 times as many establishments, and employed four times as many workers, as traditional franchising did in 2001 (Price Waterhouse Coopers 2004, p. 1).

While the United States franchising sector remains the largest in the world, franchising is increasingly a global phenomenon. Several large US-based franchisors have expanded abroad aggressively. With the development of many home-grown franchise companies, this has led to franchising sectors of many developed countries now rivalling that in the United States. According to Arthur Andersen & Co. (1995), countries such as Canada, Japan and Australia have more franchisees per capita than the United States. Still, the extent of franchising continues to vary significantly across countries.

The interest of industrial organization economists in the study of franchising emerged in the 1970s. Going back at least to Caves and Murphy (1976), Rubin (1978) and Klein (1980), economists have formulated theories about why franchising exists and why the contracts take the form they do. The economic significance of franchising in itself would easily justify this interest. However, much of the research on franchising has been carried out with a much broader goal in mind, namely to understand how firms organize their activities generally, with franchising viewed as an exemplar of the types of long-term, contract-based organizations that stand between spot market interactions and complete vertical integration, and thus a context in which to test agency and transaction cost theory. As Caves and Murphy note (1976, p. 572), 'The franchise relation raises fundamental questions concerning the nature of the firm and the extent of its integration.'

Caves and Murphy introduced many of the issues that have remained central themes in the literature, noting in particular the scale differential that gives rise to chain structures, the need to price

franchise rights to give incentives to franchisees, and the factors that lead firms to rely to varying degrees on franchising rather than company ownership. Regarding the latter, the authors emphasized the franchisor's initial need for capital, the importance of owner operators in some industry segments, and the possibility that franchisees might, through various activities and spillover effects, damage the brand. Mathewson and Winter (1985) formalized many of these ideas. Rubin (1978) pointed out the role that franchisors play in developing and maintaining the value of their brands, thereby noting explicitly that franchisor incentives also matter. Based on this idea, Bhattacharyya and Lafontaine (1995) developed a model to explain some remaining puzzling facts about the contracts, namely the degree of uniformity and stability of the financial terms in these contracts. Finally, a separate but complementary approach to explaining various aspects of franchise contracts, which focuses on self-enforcement, was proposed in part by Rubin but developed most explicitly by Klein (1980, 1995).

Perhaps what distinguishes franchising the most from other contractual contexts, however, is the amount of empirical work that has been conducted on the subject. This empirical literature has established several facts. First, it has shown that incentive issues on the franchisee's and the franchisor's side play a central role in franchise contracting (see Lafontaine and Slade 2007 for a review). It has also shown that franchisees' local profit-maximizing behaviour – or opportunism – can be a problem for franchisors. Consequently, the relationships are designed with self-enforcement in mind (see for example Brickley et al. 1991, and Kaufmann and Lafontaine 1994, on the role of contract termination and the presence of ongoing rent respectively).

In some cases, the theories and the facts have not matched so well. For example, franchising, like sharecropping, tends to be positively associated with risk (see for example Allen and Lueck 1995, on sharecropping). This is inconsistent with the typical agency argument that risk-averse agents should be insured more when the environment is more volatile. Lafontaine and Bhattacharyya (1995) and Prendergast (2002) explain this empirical 'anomaly' by noting that franchisees choose their effort level in ways that exacerbate the high and low demand signals they receive, which in turn makes the variance of outcomes – measured risk – larger for franchised than company outlets. Prendergast (2002), moreover, argues that principals will need to delegate more, and thus give higher powered incentives to agents, in uncertain environments. Ackerberg and Botticini (2002) propose instead that this anomalous effect of risk reflects an endogenous matching problem.

Finally, the literature on franchising has found that incentive requirements and mechanisms interact in important ways within a given relationship or contract (see notably Slade 1996; Bradach 1997; Brickley 1999; Lafontaine and Raynaud 2002). Moreover, competition or antitrust policy, as well as franchise-specific laws, constrain the set of contract terms franchisors can rely on. Another important – and underdeveloped – segment of the literature examines the effect of franchising on economic outcomes. The need for exogenous variation in organizational form has made this type of work difficult, but results suggest that prices, for example, are somewhat higher under franchising (see Lafontaine and Slade 2007, for a review). Much more work is needed, however, in both these areas.

See Also

- ▶ Anti-trust Enforcement
- ▶ Contract Theory
- ▶ Moral Hazard
- ▶ Principal and Agent (i)
- ▶ Principal and Agent (ii)
- ▶ Sharecropping
- ▶ Vertical Integration

Bibliography

Ackerberg, D.A., and M. Botticini. 2002. Endogenous matching and the empirical determinants of contractual form. *Journal of Political Economy* 110: 564–591.

Allen, D.W., and D. Lueck. 1995. Risk preferences and the economics of contracts. *American Economic Review* 85: 447–451.

Arthur Andersen & Co. 1995. Worldwide franchising statistics.
Bhattacharyya, S., and F. Lafontaine. 1995. Double-sided moral hazard and the nature of share contracts. *RAND Journal of Economics* 26: 761–781.
Blair, R.D., and F. Lafontaine. 2005. *The economics of franchising*. Cambridge: Cambridge University Press.
Bradach, J.L. 1997. Using the plural form in the management of restaurant chains. *Administrative Science Quarterly* 42: 276–303.
Brickley, J.A. 1999. Incentive conflicts and contracting: Evidence from franchising. *Journal of Law and Economics* 42: 745–774.
Brickley, J.A., F.H. Dark, and M.S. Weisbach. 1991. The economic effects of franchise termination laws. *Journal of Law and Economics* 34: 101–132.
Caves, R.E., and W.F. Murphy. 1976. Franchising: Firms, markets, and intangible assets. *Southern Economic Journal* 42: 572–586.
Engel, E.M.R.A., R.D. Fisher, and A. Galetovic. 2001. Least-present-value-of-revenue auctions and highway franchising. *Journal of Political Economy* 109: 993–1020.
Kaufmann, P.J., and F. Lafontaine. 1994. Costs of control: The source of economic rents for McDonald's franchisees. *Journal of Law and Economics* 37: 417–454.
Klein, B. 1980. Transaction cost determinants of 'unfair' contractual arrangements. *American Economic Review* 70: 356–362.
———. 1995. The economics of franchise contracts. *Journal of Corporate Finance* 2: 9–37.
Lafontaine, F., and S. Bhattacharyya. 1995. The role of risk in franchising. *Journal of Corporate Finance* 2: 39–74.
Lafontaine, F., and E. Raynaud. 2002. Residual claims and self enforcement as incentive mechanisms in franchise contracts: Substitutes or complements. In *The economics of contract in prospect and retrospect*, ed. E. Brousseau and J.M. Glachant. Cambridge: Cambridge University Press.
Lafontaine, F., and M.E. Slade. 2007. Vertical integration and firm boundaries: The evidence. *Journal of Economic Literature* 45(3): 631–687.
Mathewson, F., and R.A. Winter. 1985. The economics of franchise contracts. *Journal of Law and Economics* 28: 503–526.
Prager, R.A. 1990. Firm behavior in franchise monopoly markets. *RAND Journal of Economics* 21: 211–225.
Prendergast, C. 2002. The tenuous trade-off between risk and incentives. *Journal of Political Economy* 110: 1071–1102.
Price Waterhouse Coopers. 2004. Economic impact of franchised businesses.
Rubin, P. 1978. The theory of the firm and the structure of the franchise contract. *Journal of Law and Economics* 21: 223–233.
Slade, M.E. 1996. Multitask agency and contract choice: An empirical assessment. *International Economic Review* 37: 465–486.
U.S. Department of Commerce. 1988. *Franchising in the economy*, prepared by Andrew Kostecka. Washington, DC: U.S. Department of Commerce.
Zupan, M.A. 1989. Cable franchise renewals: Do incumbent firms behave opportunistically? *RAND Journal of Economics* 20: 473–482.

Franklin, Benjamin (1706–1790)

Henry W. Spiegel

One of the founding fathers of the United States, Franklin is remembered as 'the wisest American' for his many accomplishments as statesman, scientist and writer. As a writer he extolled the virtues of industry and thrift in many memorable phrases, some of which have become household maxims. They lent support to Max Weber's thesis of the Protestant origin of capitalism and were cited by him.

Franklin was a man of wide reading and pronounced intellectual curiosity, whose scientific contributions were mainly in natural science. He has, however, a number of economic writings to his credit. The two most important treat of monetary expansion and population growth. In 1728, at the age of 22, when he was active as a printer, he published *A Modest Inquiry into the Nature and Necessity of a Paper Currency*, in which he made a successful plea for an issue of colonial paper money. If money is tight, Franklin argued, interest rates will be high, prices low, immigration discouraged and imports stimulated. Moneylenders and lawyers may benefit from this, but other groups will suffer. If the paper money is issued on the security of land, the value of money will not decline.

In 1755 Franklin published *Observations Concerning the Increase of Mankind and the Peopling of Countries*. Like the *Modest Inquiry*, it was widely read and influential, and like the other work it shows the influence of Sir William Petty (1623–87). Both Petty and Franklin were convinced of the advantages of a large and swiftly growing population. While the central tendency

of Franklin's work runs counter to that of Malthus's later work on population, there are certain notions that can be found in the writings of both men: the idea that population tends to double in 25 years, and the notion of prudence as constituting a check to early marriages and thereby to population growth.

In the *Modest Inquiry* Franklin observed that 'trade in general being nothing else but the exchange of labour for labour, the value of all things is ... most justly measured by labour' (Spiegel 1960, p. 16). This elicited praise from Marx, who extolled Franklin as 'one of the first economists after William Petty who grasped the nature of value' (*Capital*, vol. 1, ch. 1). Marx also noted Franklin's definition of man as a tool-making animal, a definition he described as characteristic of Franklin's Yankeedom (ibid., ch. 11).

References

Carey, L.J. 1928. *Franklin's economic views*. New York: Doubleday.

Dorfman, J. 1946. *The economic mind in American civilization 1606–1865*, vol. 1. New York: Viking.

Spiegel, H.W. 1960. *The rise of American economic thought*. Philadelphia: Chilton.

Fraud

Edi Karni

An agent is said to have committed fraud when he misrepresents the information he has at his disposal so as to persuade another individual (principal) to choose a course of action he would not have chosen had he been properly informed. The essential element of this phenomenon is the presence of two individuals both of whom have something to gain from co-operating with each other but who have conflicting interests and differential information. More specifically, it is critical that the agent be both better informed than the principal and in a position to use his superior knowledge to affect the principal's actions so as to increase his own share of the total benefit at the principal's expense. As the choice of terminology indicates, fraud is a special case of a more general class of economic phenomena known as agency relationships. (For a more elaborate discussion and citations see Arrow 1985.)

Fraud may assume different forms. To focus our discussion, however, we consider the provision by a producer (agent) of misinformation so as to induce customers (principals) to purchase goods or services which, if adequately informed, they would not buy. Our discussion draws heavily upon Darby and Karni (1973), which was the first and so far the most elaborate attempt at an economic analysis of the phenomenon of fraud.

The Prevalence of Fraud

Fraud is as prevalent and as persistent as the asymmetrical information necessary to support it. Thus fraud may occur whenever the cost of verification of the producer's claims prior to the actual purchase of the good or service is prohibitively high. For some goods the producer's claims are easily verifiable through their use, for example, the performance of a car, the effectiveness of a painkiller. In these cases, if the population participating in the market is sufficiently stable, the scope for fraud by established firms is limited by the need to maintain their reputations. In such markets fraud may nevertheless be practised by transient firms and fly-by-night operators.

Fraudulent practices of a more persistent nature may occur in service industries where the separation of the diagnosis from provision of the service itself is impractical and where, moreover, the assessment of the quality of service is difficult if not impossible. This is the case when the ultimate performance of the good being serviced depends on several inputs and/or the relation between the service input and the ultimate performance is stochastic. To grasp the point consider a patient who complains of stomach pain. Suppose that the patient is treated with two different medications and undergoes surgery. Should the pain

disappear the patient would be unable to determine which, if any, of the possible remedies was responsible for his cure.

The Economic Consequences of Fraud

The opportunities for fraud manifest themselves in voluntary arrangements that define the principal-agent relation, whose purpose is to inhibit the actual perpetration of fraud, and in resource misallocation.

Voluntary arrangements and institutions such as formal warranties and service contracts may be regarded as insurance schemes. However, by placing responsibility for the cost of maintenance on the supplier these contracts eliminate the supplier's incentive to defraud his customers. Thus, in the absence of direct means of verification, extended warranties and service contracts may be regarded as means by which producers authenticate their claims (see Hirshleifer 1973, for a discussion of authentication as an information-induced behavioural mode). The scope for formal service contracts and warranties is limited by the usual 'moral hazard' problem. In other words, the adverse effect on the owner's incentive to take the necessary care in using the good may undermine these institutions.

A less formal arrangement is the 'client relationship'. This form of principal–agent relation is an implicit agreement that the customer will continue to patronize the service shop as long as he has no reason to suspect fraud. Lacking the means necessary for a direct assessment of the service provided, customers may exploit the opportunity afforded by repeated relations to detect whether a supplier performs at the desired level by using statistical methods. Recognizing this and the need to cultivate a clientele discourages the supplier from defrauding regular customers. This personal relationship replaces the anonymity typical of markets in which information is symmetrically endowed. Obviously this consideration does not apply to transient clientele. Indeed, large parts of the folklore surrounding the tourist industry consist of accounts of flagrant fraudulent practices. (For a more detailed discussion of the client relationship, see Karni and Darby 1973; Glazer 1984.)

The profit opportunities made possible by fraud attract resources to industries where such opportunities exist. When barriers to entry do not exist excessive profits are eliminated. The resulting resource allocation, however, is distorted as scarce resources are employed in the provision of unnecessary services.

The Deterrence of Fraud

Successful detection and prosecution of fraud have a deterrent effect that benefits society. Thus, a case can be made for social intervention. This may take the form of awarding multiple damages to successful prosecution of fraud that would reflect the full social benefit from its deterrence. Such a policy would have the effect of increasing private vigilance in dealing with fraudulent practices and, with appropriate penalties on the practitioners, reduce the amount of fraud to a socially desirable level. Alternatively, adherence to non-fraudulent practices may be enforced by the law enforcement agencies of the government. (For a detailed discussion, see Darby and Karni 1973.)

Since the provision of misinformation may just as well be the result of sheer incompetence as of intentional deception, successful fraud-deterring policy will also increase the competence level of the suppliers of services. Unlike the elimination of intentional misrepresentation of information, however, increasing the level of competence involves investment of scarce resources on the part of the suppliers. Therefore, in setting the goals for a policy whose aim is to reduce fraud, the additional gains from the associated increase in the level of competence must be weighed against the corresponding resource cost. The optimal level of fraud may not be zero.

See Also

► Asymmetric Information

Bibliography

Arrow, K.J. 1985. The economics of agency. In *Principals and agents: The structure of business*, ed. J.N. Pratt and R. Zeckhauser. Cambridge, MA: Harvard Business School Press.

Darby, M.R., and E. Karni. 1973. Free competition and the optimal amount of fraud. *Journal of Law and Economics* 16(1): 67–88.

Glazer, A. 1984. The client relationship and a 'just' price. *American Economic Review* 74(5): 1089–95.

Hirshleifer, J. 1973. Where are we in the theory of information? *American Economic Review* 63(2): 31–9.

Free Banking

C. F. Dunbar

Free banking is the term applied in the United States to a system under which (1) banking powers are granted to all applicants under certain prescribed conditions, and (2) bank-notes issued under such authority are protected by a deposit of security held by the government which establishes the system. The earlier banks in the United States, whether established by congress or by the state legislatures, were organized under special charters. Various expedients were resorted to for the prevention of unsound issues, with various degrees of success, but without arriving at any generally acceptable method. The suspension of specie payments in May 1837, and the extraordinary confusion of the paper currency which ensued, finally brought the general discontent to a climax in New York, and the legislature of that state, in June 1838, passed an act for the free organization of banks issuing a secured currency. Under this act, as amended and revised, any group of persons proposing to form a banking association, and contributing a capital in no case less than $25,000, say £5000, can be incorporated with full banking powers, subject to uniform regulations as to the conduct of their business, its supervision by the state, and their corporate liabilities and duties. Individual bankers and firms, who use the name 'bank', are also required to conform to the system, although they may remain unincorporated. The right of issue is given to any association or individual coming under the system. The notes are prepared and registered by a public officer, are delivered to the issuing bank only after the deposit of security of a prescribed kind and amount, and must be signed by the officers of the bank before issue. Banks organized upon such a system are called free banks.

Free banking does not imply, then, an unrestricted management of the business, or complete liberty in the issue of notes. Such a system is called free because the right to organize, upon compliance with fixed conditions, is extended to all, free from any requirement of special legislation. It is not essential that there should be any engagement by the state to make the notes good, if the security, of which the state is trustee, proves insufficient. Neither does the deposit of security for the ultimate payment of the notes answer the question as to proper provision for daily redemption. As the provision for secured notes gave promise of insuring the ultimate solvency of bank notes, it settled the one banking question as to which the public were most sensitive, and enabled the legislature to renounce the task of deciding upon applications for special charters. The system adopted by New York was copied by many other states before the civil war, but in some cases with relaxations which impaired its safety. In 1861 the New York free banks, having on deposit stocks of the United States and other solid securities, met the strain of war with success. In several states, where the law was less rigid, many free banks went down, and their notes, secured in some cases chiefly by bonds of seceded states and others in low credit, caused heavy losses to the holders. Two years later Congress adopted the free banking system on a great scale, by a law providing for national banks, to be organized on application under a general act, and to issue notes with United States bonds as the only admissible security. In 1865 Congress laid a tax of ten per cent on all bank notes other than national, thus excluding from the field all issues authorized by the states. Several of the states, however, still retain their laws as to circulation, although these have been entirely dormant since 1866. Free banking under the national system was for some years seriously limited, by the

provision that the aggregate of notes issued by all the national banks should not exceed $300,000,000, afterwards $354,000,000 (say £60,000,000, and £70,800,000) although the organization of banks was still free to all. The act of 1875, for resuming specie payments, removed the limit of aggregate circulation, and thus completely established free banking under the national government. The rapid rise in price of United States bonds, and the low return yielded by an investment in them, have since put a new check upon the system; and if the use of bank-notes is to continue, the alternative may soon be presented, of either finding for deposit by national banks some other security than United States bonds, or removing the prohibitory tax upon issues authorized by the states.

Free Banking Era

Arthur J. Rolnick and Warren E. Weber

Abstract

In the free banking era entry into banking was virtually unrestrained, banks could issue their own currency and governments did not insure banks; many banks closed and many noteholders reportedly suffered. An early view of this period is that free entry led to banks over-issuing notes, resulting in large losses for noteholders. More recent research has shown that this is incorrect. Although such failures and losses did occur, these were generally due to the capital losses banks suffered when the prices of the state bonds backing their notes fell, rather than to note over-issuance or fraudulent banking practices.

Keywords

Free banking; Free banking era; Free banking laws; Wildcat banks

JEL Classifications

G2

Imagine the US economy without Federal Reserve notes, that is, without a uniform currency. Instead, imagine that the currency consists of notes issued by privately owned banks and that are redeemable in specie on demand. And imagine that to enter the banking business is relatively easy, so that the notes of hundreds of banks exist. And imagine as you travel around the country, notes of out-of-town banks are not readily accepted as means of payment at par because the solvency of such banks is difficult to ascertain.

How well would such a banking system function? In particular, with free entry into banking, would banks not have an incentive to over-issue their notes, leaving the public holding worthless pieces of paper when the banks failed? And would trade not be difficult without the existence of a uniform currency? Indeed, a reading of historical accounts of the so-called free banking era – the 26 years from 1837 to 1863, a period when entry into banking was relatively free and banks issued their own notes – would lead to this conclusion. The prevailing view of this period, at least until the mid-1970s, was that allowing such freedom in banking was a mistake. However, a more recent examination of the era reveals that while the free banking system was not without its problems, free banks and their noteholders fared much better than has often been portrayed.

The Beginning of Free Banking

Prior to 1837, to establish a bank in the United States was a very cumbersome, and at times political, process. Individuals who wanted to start a bank had to obtain a charter from the legislature of the state in which they wanted to operate. Beginning in 1837, some states reformed their bank-chartering systems so that entry into the banking industry would be easier. States tempered the goal of easy entry with another goal: to provide the public with a safe bank currency. Most states attempted to reach these goals by enacting what were called *free banking laws*.

The first free banking law was proposed in New York. Its provisions openly aimed at both easy entry and safety. The law allowed anyone to

operate a bank as long as two basic requirements were met: (*a*) all notes the bank issued had to be backed by state bonds deposited at the state auditor's office and (b) all notes had to be redeemable on demand at par, or face, value. If the bank failed to redeem notes presented for payment, however, the auditor would close the bank, sell the bonds, and pay off the noteholders. If the bond sale did not generate enough specie to redeem the bank's notes at par, noteholders had additional protection by having first legal claim to the bank's other assets. Thus, free banking meant free *entry* into banking; it did not mean laissez-faire banking.

New York's proposed free banking law became the basic blueprint for the free banking laws in other states. (Michigan actually passed a free banking law modelled on the New York proposal a year before the legislation was passed there.) Table 1 shows which states passed free banking laws and when the laws passed. Note that of the states that passed such legislation, most did so in the 1850s.

Free Banking Era, Table 1 US states with and without free banking laws by 1860

States with free banking laws	Year law passed	States without free banking laws
Michigan	1837[a]	Arkansas
Georgia	1838[b]	California
New York	1838	Delaware
Alabama	1849[b]	Kentucky
New Jersey	1850	Maine
Illinois	1851	Maryland
Massachusetts	1851[b]	Mississippi
Ohio	1851[c]	Missouri
Vermont	1851[b]	New Hampshire
Connecticut	1852	North Carolina
Indiana	1852	Oregon
Tennessee	1852[b]	Rhode Island
Wisconsin	1852	South Carolina
Florida	1853[b]	Texas
Louisiana	1853	Virginia
Iowa	1858[b]	
Minnesota	1858	
Pennsylvania	1860[b]	

[a]Michigan prohibited free banking after 1839 and then passed a new free banking law in 1857
[b]According to Rockoff, very little free banking was done under the laws of these states
[c]In 1845, Ohio passed a law that provided for the establishment of 'independent banks' with a bond- secured note issue
Source: Rockoff (1975, pp. 3, 125–30)

The Experience

One effect of the free banking laws was to increase the number of banks. In Michigan, for example, the number of banks rose from ten before the law was passed in March 1837 to 33 one year later. In New York the number of banks rose from 97 before the law was passed in March 1838 to 162 three years later. And Indiana, Illinois, and Wisconsin, which each had only one bank in existence when their free banking laws were passed, saw 13, 41, and 15 new banks established respectively within two years. Minnesota had no banks when its free banking law was passed; it saw 16 banks established within one year. In total, of the almost 2,300 banks that existed in the United States prior to the Civil War, slightly more than three-eighths were established or operated under a free banking law (Weber 2006).

Free banking, however, must also be judged by the laws' second objective – by how many banks survived and provided their communities with a stable source of banking services, especially a safe currency. Measured by this criterion, free banking is generally considered a failure.

Michigan's disastrous experience with free banking is probably the most famous. By the end of 1839, less than two years after its free banking law was passed, all but four of Michigan's free banks closed (Rockoff 1975, p. 96).

Although explicit loss data do not exist, it has been estimated that the total loss to Michigan's noteholders was as high as four million dollars. This would have been nearly 45 per cent of Michigan's annual income in 1840 (Rockoff 1975, pp. 17–48). Other states' experiences with free banking, while not as famous as Michigan's, were almost as bad. Of the 16 free banks that opened under Minnesota's 1858 law, for example, 11 closed by 1863. And many that closed left their noteholders with very little.

However, some states had positive experiences with free banking. New York had very few free bank failures and noteholder losses after 1843.

Indiana had much the same record after 1854. And all the failures and losses experienced by Wisconsin free banks occurred in 1861 after the Civil War had begun and the bonds issued by Southern states had greatly depreciated in value.

Free Banking was not Wildcat Banking

According to some historians and economists writing about this period (see, for example, Hammond 1985, p. 618; Knox 1903, p. 747; and Luckett 1980, p. 242), the losses experienced under free banking were due to fraudulent banking practices by so-called wildcat banks. These were banks that purportedly located redemption offices in remote areas, issued notes far in excess of what they planned to redeem, and then disappeared, leaving the public with notes worth considerably less than their original value.

Although some wildcat banking may have occurred, this explanation is not appropriate for most free banking experience because the data do not support it. Wildcat banks supposedly stayed in business for only a few months, after which time their noteholders sustained losses. However, in New York, Indiana, Wisconsin, and Minnesota – four states that were supposed to have had many wildcats – free banks were generally not short-lived.

Most losses to the holders of free bank notes were due not to fraud, but to capital losses suffered by the banks because of several substantial drops in the prices of the state bonds that were required to back the notes they issued. Moreover, while these declines in bond prices may have been induced by any number of economic developments, they were not induced by wildcat banks.

Summary and Conclusion

The free banking era was a time when entry into banking was virtually unrestrained, when banks could issue their own currency and when the government did not insure banks. It was also a time when many banks closed and many noteholders reportedly suffered. An early view of this period is that free entry led to banks overissuing notes, resulting in large losses for noteholders. More recent research has shown that this view is not correct. Although free bank failures and noteholder losses did occur, these were generally due to capital losses banks suffered when the prices of the state bonds backing their notes fell. In general, they were not due to note over-issuance or fraudulent banking practices.

See Also

- ▶ Banking Crises
- ▶ Banking School, Currency School, Free Banking School
- ▶ Monetary Economics, History of

Bibliography

Hammond, B. 1985. *Banks and politics in America*. Princeton: Princeton University Press.

Knox, J.J. 1903. *A history of banking in the United States*. New York: Bradford Rhodes.

Luckett, D.G. 1980. *Money and banking*. New York: McGraw-Hill.

Rockoff, H. 1974. The free banking era: A reexamination. *Journal of Money, Credit and Banking* 6: 141–167.

Rockoff, H. 1975. *The free banking era: A re-examination*. Dissertations in American History. New York: Arno Press. Ph.D. dissertation, University of Chicago, 1972.

Rolnick, A.J., and W.E. Weber. 1983. New evidence on the free banking era. *American Economic Review* 73: 1080–1091.

Rolnick, A.J., and W.E. Weber. 1984. The causes of free bank failures: A detailed examination. *Journal of Monetary Economics* 14: 267–291.

Weber, W.E. 2006. Early state banks in the United States: How many were there and when did they exist? *Journal of Economic History* 66: 433–455.

Free Disposal

Theodore C. Bergstrom

Keywords

Competitive equilibrium; Excess demand; Free disposal; Kakutani's fixed point theorem; Monotonicity

JEL Classifications
D5

> 'I should like to buy an egg, please' she said timidly. 'How do you sell them?' 'Fivepence farthing for one – twopence for two,' the Sheep replied.
> 'Then two are cheaper than one?' Alice said, taking out her purse.
> 'Only you must eat them both if you buy two,' said the Sheep.
> 'Then I'll have one please', said Alice, as she put the money down on the counter. For she thought to herself, 'They mightn't be at all nice, you know.'
> (Lewis Carroll, *Through the Looking-Glass*)

If I dislike a commodity, you may have to pay to get me to accept it. But so long as some otherwise non-sated consumer finds this commodity to be desirable, or at least harmless, it could not have a negative price in competitive equilibrium. Likewise, if some firm can dispose of arbitrary amounts of a commodity without using any other inputs or producing any other (possibly noxious) outputs, its price in competitive equilibrium cannot be negative. Therefore competitive equilibrium analysis can be confined to the case of non-negative prices if every commodity is either harmless to someone or freely disposable.

If a commodity is not freely disposable and is a 'bad' in the sense that everyone prefers less of it to more, it is possible to redefine the 'commodity' as the absence of the bad. The commodity so defined can then be treated as a good with a positive price. More generally, it might be possible to choose some alternative coordinate system in which to measure commodity bundles so that in the new coordinate system either there is free disposability or more is preferred to less. But if people are willing to pay a positive sum for a small amount of a commodity and less for a large amount, then the question of whether that commodity will have a positive or negative price in competitive equilibrium cannot be decided in advance. The sign of the equilibrium price will in general depend on supplies of this and other goods and on the detailed configuration of preferences in the economy.

Sometimes a noxious by-product of production or consumption can be transformed into a useful output if sufficient other resources are used. Then the equilibrium price for the by-product may be either positive or negative, depending on the prices of the other inputs and of the output into which it is transformed. This is particularly evident when commodities are distinguished by location. Garbage located in the centre of a city is undesirable to everyone. To bury or incinerate it is costly and generates no valuable outputs. Therefore, if garbage is disposed of in this way, its equilibrium price must be negative. But the garbage could be transported to the country, boiled and fed to pigs. Depending on the costs of this process and the price of pork, it may turn out that converting garbage to pig feed is profitable even when garbage at the city centre has a zero or positive price. Both the ultimate disposition of garbage and the sign of its price have to be determined endogenously in the competitive process.

Early proofs of the existence of competitive equilibrium (Arrow and Debreu 1954; Gale 1955; Debreu 1959) assumed that all commodities are freely disposable or, equivalently, defined equilibrium so as to allow the possibility that in equilibrium some goods might be in excess supply but have zero price. Debreu (1956) shows how the assumptions of free disposal and monotonicity can be greatly relaxed. McKenzie (1959) and Debreu (1962) present general theorems on the existence of equilibrium in which free disposal is not assumed. Rader (1972), Hart and Kuhn (1975), Bergstrom (1976) and Shafer (1976) suggest further generalizations and simplifications in dealing with negative prices in equilibrium.

The formal treatment of negative prices in existence proofs presents an interesting mathematical problem. Most of the standard existence proofs apply the Kakutani fixed-point theorem to a correspondence that maps the set of possible equilibrium prices into itself in such a way that a fixed point for the mapping is a competitive equilibrium price vector. The Kakutani theorem applies to an upper hemicontinuous mapping from a closed bounded convex set to its compact, convex subsets. If the only prices to be considered are non-negative, then the domain for this correspondence can be chosen to be the unit simplex. If all price vectors, positive and negative, must be considered, then an obvious candidate for the domain

of this mapping would be the unit sphere $\{p \in R^n | \ p \cdot p = 1\}$. But this is not a convex set. The closed unit ball $\{p \in R^n | \ p \cdot p \leq 1\}$ is a convex set, but it contains the vector zero, at which point the excess demand mapping is not upper hemicontinuous.

Debreu (1956) solved this problem neatly in a brief, elegant paper that has received less attention than it deserves. The existence proofs that assume free disposability of all goods had shown that there exists a non-negative price vector at which the excess demand vector is either zero or belongs to the negative orthant. Debreu generalized this result to show that if there is free disposability on any convex cone which is not a linear subspace, then a price vector can be found at which excess demand is either zero or belongs to the cone of free disposability. Furthermore, this price vector gives a non-positive value to every activity in the cone of free disposability. In particular, consider the case where one good is assumed to be freely disposable. Then, from Debreu's theorem, it follows that there exists some price vector at which excess demand for all goods other than the freely disposable good is zero, and at which there is either zero or negative excess demand for the freely disposable good. From Walras's Law and the fact that that there exists some price vector at which excess demand for all other goods is zero, it follows that the price of the freely disposable good can be positive only if excess demand is zero. Therefore this price vector is a competitive equilibrium. Thus Debreu weakened the free disposability assumption from 'all goods are freely disposable' to 'at least one good is freely disposable'.

We can take Debreu's argument one step further and eliminate the assumption of even one freely disposable good. Nowhere in Debreu's proof is it necessary to assume that the freely disposable good is desirable to anyone. This suggests that the existence of a freely disposable good is not likely to be essential for the existence of equilibrium. For suppose that there is an economy with no freely disposable goods. A fictional good could be introduced which is freely disposable but totally useless and totally harmless to everyone. For the augmented economy found by adding this fictional good to the original economy, by Debreu's theorem there would exist a competitive equilibrium. In this new economy it turns out that the equilibrium price of the useless, freely disposable good must be zero and the vector of equilibrium prices for the other goods can serve as a competitive equilibrium price vector for the original economy.

The approach taken by Bergstrom (1976) is equivalent to introducing a useless and harmless fictional good into Debreu's model. Taking the formal steps of this argument directly without intermediary fictions leads to an upper hemicontinuous mapping from the unit ball into itself for which there is a fixed point on the boundary of the unit ball. This fixed point turns out to be a competitive equilibrium price vector. An interesting alternative approach was taken by Rader (1972) and by Hart and Kuhn (1975). Instead of the Kakutani theorem, they use a theorem about fixed and antipodal points of a continuous mapping from the unit sphere into itself, and are thereby able to deal with all prices on the unit sphere as potential equilibrium prices.

The first and second welfare theorems and the theorems about the equivalence between the core and the set of competitive equilibria apply straightforwardly when there is not free disposal. For example, in order to prove the Pareto optimality of competitive equilibrium in an exchange economy, we simply argue along the usual lines that if any allocation is Pareto superior to a competitive equilibrium, then at the original competitive prices, the aggregate value of consumption in the proposed Pareto superior allocation must exceed the aggregate value of initial endowments. But if the proposed allocation is feasible, then the aggregate consumption vector in the proposed allocation must equal the aggregate initial endowment vector. It follows, whether prices are positive, negative or zero that if the two vectors are equal they must have the same value at the competitive price vector. Therefore there cannot be a feasible allocation which is Pareto superior to a competitive equilibrium. Similar arguments apply to the core theorem. The only matter in which a bit of care must be taken is in defining the activities available to a potential blocking coalition so as to exclude the possibility of dumping undesirable commodities. This simply

amounts to the assumption that a blocking coalition must exactly equalize its total consumption of all goods to its total endowment.

See Also

- ▶ Fixed Point Theorems
- ▶ General Equilibrium

Bibliography

Arrow, K.J., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Bergstrom, T. 1976. How to discard free disposability – At no cost. *Journal of Mathematical Economics* 3: 131–134.

Debreu, G. 1956. Market equilibrium. *Proceedings of the National Academy of Sciences of the USA* 42: 876–878.

Debreu, G. 1959. *The theory of value*. New York: Wiley.

Debreu, G. 1962. New concepts and techniques for equilibrium analysis. *International Economic Review* 3: 257–273.

Gale, D. 1955. The law of supply and demand. *Mathematica Scandinavica* 3: 155–169.

Hart, O., and H. Kuhn. 1975. A proof of the existence of equilibrium without the free disposal assumption. *Journal of Mathematical Economics* 2: 335–343.

McKenzie, L. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27: 54–71.

Rader, T. 1972. *Theory of general economic equilibrium*. New York: Academic Press.

Shafer, W. 1976. Equilibrium in economies without ordered preferences or free disposal. *Journal of Mathematical Economics* 3: 135–137.

Free Goods

Ian Steedman

Free goods are 'goods', whether consumer goods or productive inputs, which are useful but not scarce; they are in sufficiently abundant supply that all agents can have as much of them as they wish at zero social opportunity costs (cf. ch. 11, §3, of Carl Menger's *Principles of Economics*, 1871). Goods which have a positive social opportunity cost but a zero price – for example, because there are no property rights in them, or because they are fully subsidized – are *not* free goods. Any 'gift of nature', whether it be a good such as air, or a primary input such as labour or land (in the narrow sense), might be a free good under certain circumstances. But a produced commodity can be a free good, other than in the market period, only if it is a joint product. As is at once obvious from the example of air, the free nature of a good is not an intrinsic property; thus air above the earth's surface is, in most circumstances, a free good but air under water or in deep mines is not. More abstractly, then, a free good is a good for which supply is not less than demand at a zero price (in the sense of social opportunity cost). But since both supply of and demand for any good depend on the prices of all goods, it is clear that whether a particular good is or is not a free good is a general equilibrium, not a partial equilibrium, issue.

Consider first a Walrasian analysis of general equilibrium. Under the standard assumptions of such an analysis, Walras's Law (or identity) holds, so that $pS \equiv 0$, *where* p is a row vector of prices and S a column vector of excess supplies. (This is an *identity*, holding at all prices, not only at equilibrium prices.) Now, by definition, in a Walrasian analysis any equilibrium excess supply vector satisfies $S^* \geq 0$. Hence if it is ensured that any equilibrium price vector satisfies $p^* \geq 0$, it follows – from $pS \equiv 0$ and $S^* \geq 0$ – that if $S_j^* > 0$ then $p_j^* = 0$. That is, the Rule of Free Goods holds in such a Walrasian equilibrium, applying to all 'goods', whether produced or non-produced. Two points are to be noted. The less important one is that while $S_j^* > 0$ implies $p_j^* = 0$, $p_j^* = 0$ does not imply $S_j^* > 0$, since $p_j^* = 0 = S_j^*$ is possible. The more important point is that the Rule of Free Goods is not implied by $pS \equiv 0$ and $S^* \geq 0$ alone; they must be supported by the condition $p* \geq 0$. This last condition is often underpinned by an assumption of the possibility *free disposal* (see below). Such an assumption rules out the possibility that any $p_j^* < 0$, for there would be an unlimited demand for a good for which one 'paid' a negative price – that is *received* a positive price – and which one could dispose of at zero cost.

It was noted above that the Rule of Free Goods is applied in Walrasian flex-price analyses to both products and primary inputs. With respect to the latter, it is instructive to consider the linear programming formulation which is sometimes given for the 'supply' side of a general equilibrium existence proof for an economy with linear technical conditions. In the primal problem one is asked to maximize the value of net output, at parametrically given product prices, subject to not using more than the exogenously fixed supply of any primary input. The complementary slackness conditions, corresponding to these last constraints, give immediate expression to the Rule of Free Goods, as applied to the primary inputs. And the non-negativity constraints in the dual problem stipulate, of course, that the solution factor prices cannot be negative. Thus every solution factor price will be non-negative and will be zero if the relevant factor is less than fully utilized.

It is essential to note that not all types of economic analysis impose the Rule of Free Goods with respect to all primary inputs. In the von Neumann model, for example, that rule is certainly imposed with respect to all the produced commodities, but it is not applied to labour, which receives an exogenously given real wage bundle which is independent of the degree of utilization of labour. At most, one could say that a 'Rule of Zero "Excess" Wages' is applied because labour is less than fully employed. Similarly, in Keynes's analysis the presence of involuntarily unemployed labour does not drive the wage to zero but only to an exogenously given minimum (a market level reservation price). Clearly, then, the three assertions, $pS \equiv 0$, $S^* \geq 0$ and $p^* \geq 0$ are not all acepted within Keynes's analysis. But since $S^* \geq 0$ and $p^* \geq 0$ *are* accepted, it can only be Walras's Law which is being rejected – and this is indeed the case, for in Keynes's analysis we have only the condition that, the elements of S being defined in terms of *desired* supplies and demands, $pS \geq 0$. The weaker relation is, of course, perfectly consistent with $S_j^* \geq 0$ and $p_j^* \geq 0$ (see Morishima, 1976, pp. 203–11).

It was noted above that the 'free-disposal' assumption has the convenient consequence that no equilibrium price can be negative; this means that the search for Walrasian equilibrium price vectors can be confined to the unit simplex. Although this restriction on prices is *not* a necessary ingredient of all general equilibrium existence proofs, the free-disposal assumption is sufficiently widely adopted (it is sometimes even described as obviously reasonable) to merit a close examination of its justification. Consider first the proposal that the commodity to be disposed of in a disposal process is the *only* input to the latter. This means that the only form of 'disposal activity' allowed is that of simply *leaving* the commodity to be disposed of *where* it is and leaving it *as* it is. If it moves or changes its form, that must be the result solely of non-human and non-produced agencies. (Note that one cannot defend the disposal activity assumption by saying that it applies only to the 'last stage' of a real-world-like disposal process, which first uses labour, lorries, etc. to take waste chemicals, for example, to a particular place. This is because the assumption is supposed to apply to *all* commodities, including, for example, the chemical waste *situated at the point of its production*).

This leads us naturally to a consideration of the second and even more objectionable – aspect of the disposal activity assumption, the proposal that the activity has *no* outputs. Taken literally, this proposal simply contradicts one of *the* most fundamental laws constituting our conception of the physical universe – the law of conservation of mass-energy. If one takes the conservation law for granted, for the purposes of economic theory, then *either* the zero-output assumption is incomprehensible *or* it means that all the outputs from the disposal process lie outside the commodity set which is taken as the basis for the economic analysis. A defence of the latter interpretation would have to involve both an account of the principles according to which that set is defined on a non-arbitrary basis and an explanation of why the outputs of disposal processes can be supposed – non-arbitrarily – to lie outside that set.

It might be said that the disposal-activity assumption simply provides one interpretation of the basic axiom of free disposal $x \in X$ and $x' \leq x$ implies $x' \in X$, where X is the production set) and that the latter may be acceptable even while the

former is not. How then might the axiom be understood in the absence of disposal processes? Suppose that together with all the other inputs and outputs (which will be held constant), a certain input of fertilizer and a certain output of maize define an activity belonging to the production set. It is then proposed that, *ceteris paribus*, the same fertilizer input and a smaller maize output also define a feasible activity. We cannot suppose that some of the fertilizer is simply not used, for then an output (that of fertilizer) would have been increased. Thus all the fertilizer must be used. If it is used in the same way as 'before' then a smaller maize output, *ceteris paribus*, involves different laws of nature. If it is used but used differently from 'before', then some other input has changed, contrary to hypothesis. Hence the presence of a disposal activity is, after all, required.

In the above example, the 'other input' which has changed when fertilizer is used differently is some human agency. For to say that fertilizer is used differently is precisely to say that *someone* has acted differently. Suppose then that we now change the example, replacing the given fertilizer input by a given quantity of a specified type of labour input and including fertilizer amongst the (given) 'other' inputs and outputs. In the absence of disposal processes, does the fact that a certain labour input and a certain maize output define an activity in the production set, mean that the *same* labour input and a *smaller* mazie output (perhaps even a *negative* one) also define such an activity? If free disposal is ruled out, the laws of nature are constant and labour is precisely defined, the answer would again seem to be No. Thus, again, the free-disposal axiom does indeed rest on the presence of disposal activities. Objections to the disposal-activity assumption are thus also objections to the axiom of free disposal itself.

It has already been noted that general equilibrium existence proofs can dispense with the free-disposal axiom and that Keynes's theory does not apply the Rule of Free Goods to labour. More generally, the rule of free goods should not simply be assumed to apply to non-produced inputs, for it must always be considered whether their owners place a positive reservation price on them. With respect to produced commodities, free disposal should not be assumed (for the reasons given above), as it commonly is in linear programming models, in studies of balanced growth within closed production models with convex cone production sets, and in proofs of turnpike theorems. In each case, on dispensing with the free-disposal axiom, one must decide how to represent preferences over 'bads'. These apparently abstract issues are, of course, of immediate relevance in the discussion of such policy issues as pollution control, environmental protection and waste disposal. (If there were no joint production, or if free disposal were possible, there could be no problems of pollution control and waste disposal.) When there are disposal activities which involve a negligible private cost but a significant social cost, policy will involve bringing the positive costs of disposal to bear on the individual agents concerned. This may induce them, in turn, to discover or invent new uses for the previously undesired 'commodities'; the costly nature of disposal has spurred changes in technical knowledge.

See Also

- ▶ Free Disposal
- ▶ General Equilibrium

Bibliography

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*. New York: Wiley.

Morishima, M. 1976. *The economic theory of modern society*. Cambridge: Cambridge University Press.

Free Lunch

Robert Hessen

'There's no such thing as a free lunch' dates back to the 19th century, when saloon and tavern owners advertised 'free' sandwiches and titbits

to attract mid-day patrons. Anyone who ate without buying a beverage soon discovered that 'free lunch' wasn't meant to be taken literally; he would be tossed out unceremoniously.

'Free lunch' passed over into political economy during the New Deal era, and is loosely credited to various conservative journalists, including H.L. Mencken, Albert Jay Nock, Henry Hazlitt, Frank Chodorov and Isabel Paterson. (All efforts to identify the true originator proved unavailing.) The phrase signified that the welfare state is an illusion: government possesses no wealth of its own, so it can only redistribute wealth it has seized by taxation.

During the Vietnam war era, 'free lunch' took on a libertarian cast. When defenders of the draft argued that young men *owed* military service because they had accepted free tuition and subsidized school lunches as youngsters, the 'free lunch' expression became a libertarian shorthand to denote that citizens never get something for nothing, that sooner or later they are presented with a bill for all the favours or 'freebies' they accepted from government.

'Free lunch' would have passed into oblivion if it had not been able to pass a crucial test of its viability in the marketplace of ideas. In the early 1970s, every political or philosophical idea had to be able to fit on a T-shirt or automobile bumpersticker. The new version, TANSTAAFL (there ain't no such thing as a free lunch), was popularized in a science fiction bestseller by Robert Heinlein (*The Moon is a Harsh Mistress*) and in Milton Friedman's widely read columns in *Newsweek* magazine.

Free Trade and Protection

R. Findlay

The question of 'free trade *versus* protection' is one of the oldest and most controversial issues in economics. The present article will not attempt to review this controversy from the standpoint of the history of doctrine, nor will it attempt to trace the evolution of trade policy in particular countries. Its focus is on the analytic aspects of the problem as discussed in the modern literature, which can be taken as dating from the seminal investigations of Samuelson (1939, 1962), in the tradition of Paretian welfare economics.

In keeping with this tradition we postulate that the criterion in terms of which any economic situation is to be evaluated, the 'social welfare function', is of the 'individualistic' type, i.e. it depends only upon the well-being of the individual agents themselves in terms of their own preferences, rather than on objectives such as national self-sufficiency, economic growth or some other vaguely defined concept of national interest.

An essential distinction to bear in mind is whether a 'cosmopolitan' or 'nationalist' perspective is adopted, i.e. are all individuals wherever located to count or only those belonging to the 'home' country. The modern view is that free trade is Pareto-optimal in the first case but not in the second, if the home country possesses some degree of monopoly power in foreign trade, which it can then exploit to improve the welfare of its own nationals at the expense of foreigners.

The first result follows from the familiar proposition that a perfectly competitive equilibrium is Pareto-optimal, in the absence of externalities in production and consumption. Marginal rates of substitution in consumption are equal for all individuals everywhere, since they face the same relative prices under free trade. Marginal rates of transformation in production are also everywhere equal, for the same reason, and equal to the corresponding marginal rates of substitution in consumption. The necessary conditions for a Pareto-optimum are therefore satisfied, and sufficiency can be shown to follow from the convexity of preference and production sets.

Since free trade is therefore globally Pareto-optimal, any restriction of trade such as a tariff or quota must be at the expense of someone. The idea of 'letting the foreigner pay the duty' is at the heart of what is known as the 'optimum tariff' argument. If we adopt a purely 'nationalist' perspective, ignoring the effects of our actions on foreign welfare, it is possible to raise our own welfare by a suitable degree of trade restriction to take advantage of our

monopoly power in international markets. If the home country is 'small', in the sense that it faces fixed relative prices in the world market for all tradable goods, any tariff or import quota that it adopts will simply reduce its volume of trade without improving its terms of trade, so that free trade is the first-best policy even with a nationalist perspective. If it does have monopoly power, however, it can balance the improvement in the terms of trade resulting from its restrictive policy against the reduction in the volume of trade that this entails. It can be shown that the formula for this 'optimum tariff' is equal to the reciprocal of the foreign elasticity of demand for imports minus one. The marginal cost of imports, and the marginal revenue for exports, deviate from world prices in the presence of monopoly power by the home country. This is why domestic producers and consumers have to equate their marginal rates of transformation and substitution to tariff-inclusive domestic prices rather than the world prices that would prevail under free trade.

The optimum tariff argument, going back at least to J.S. Mill but re-stated in modern terms of Bickerdike, Kaldor, Samuelson, Graaff and others, is the *only* argument for national trade restriction or 'protection' of import-competing sectors that the modern theory of trade and welfare recognizes. Even then, the argument that a tariff increases national welfare only holds strictly if it is assumed that foreigners do not retaliate, setting off a 'tariff war'. The outcome of such a process, as Johnson (1954) showed, is uncertain, with everybody worse off than under free trade a distinct possibility. Trade policy at the regional or global level thus becomes another example of the familiar 'prisoners' dilemma' situation explored in game theory.

The famous 'infant industry' argument for protection, on the grounds that it takes time for the arts of manufacture to be learnt, thus justifying temporary assistance in the form of tariffs imposed on competing imports, is *not* accepted as a legitimate 'first best' argument for tariffs by the modern theory. The reason is that even if manufacturing production creates externalities in the training of labour and the formation of skills the 'first best' intervention would be an output subsidy rather than a tariff. The reason is that the output subsidy would have the same beneficial effects on learning as the tariff but *without* the restrictive effect on imports and consumption. Welfare would therefore be higher. Similarly, the argument that tariff protection is necessary to tide over initial losses is countered by the contention that this could be accomplished by the capital market, any imperfections of which are best dealt with directly. Other arguments for tariffs, on the ground that urban wages are artificially high compared with rural wages, thus requiring off-setting tariff protection for manufactures, are also countered by the argument that an urban wage subsidy is the best intervention in this case.

All these separate cases are covered by the powerful and elegant theory of optimal intervention, developed by Bhagwati and Ramaswami (1963), and extended by Johnson (1965), Bhagwati (1971) and Corden (1974) with important earlier contributions by Haberler (1950) and Hagen (1958). The basic principle is that if a perfectly competitive equilibrium is not Pareto-optimal from a national perspective, it must be because there is some 'distortion' (see article) in international or domestic product and factor markets. The optimal intervention is to eliminate the distortion 'at the source', rather than to attempt to off-set it by an intervention that creates some other distortion as well. Thus, in keeping with the Lipsey-Lancaster theory of the 'second best', tariffs *may* improve national welfare in all of the above cases, but it is only in the 'optimum tariff' case that they constitute a 'first best' intervention.

The theory of optimal intervention, however, assumes that the subsidies necessary to maximize national welfare can be financed by non-distortionary means, such as lump sum taxes, and that there are no collection and disbursement costs. If these are allowed for, and it is recognized that any means of finance is itself going to be distortionary, the case for tariffs as 'second best' instruments will presumably become stronger, relative to the output and wage subsidies that the theory of optimal intervention blithely dispenses in disregard of any realistic government budget constraint.

See Also

- ▶ Heckscher–Ohlin Trade Theory
- ▶ International Trade
- ▶ Tariffs
- ▶ Terms of Trade

Bibliography

Bhagwati, J.N. 1971. The generalized theory of distortions and welfare. In *Trade, balance of payments and growth*, ed. J.N. Bhagwati et al. Amsterdam: North-Holland.

Bhagwati, J.N., and V.K. Ramaswami. 1963. Domestic distortions, tariffs and the theory of optimum subsidy. *Journal of Political Economy* 71: 44–50.

Corden, W.M. 1974. *Trade policy and economic welfare*. Oxford: Oxford University Press.

Haberler, G. 1950. Some problems in the pure theory of international trade. *Economic Journal* 60: 223–240.

Hagen, E.E. 1958. An economic justification of protectionism. *Quarterly Journal of Economics* 72: 496–514.

Johnson, H.G. 1954. Optimum tariffs and retaliation. *Review of Economic Studies* 21: 142–153.

Johnson, H.G. 1965. Optimal trade intervention in the presence of domestic distortions. In *Trade, growth and the balance of payments*, ed. R.E. Caves et al. Chicago: Rand-McNally.

Samuelson, P.A. 1939. The gains from international trade. *Canadian Journal of Economics and Political Science* 5: 195–205.

Samuelson, P.A. 1962. The gains from international trade once again. *Economic Journal* 72: 820–829.

Friedman, Milton (1912–2006)

Alan Walters

Abstract

Milton Friedman is widely regarded as one of the most important economists of the 20th century. He is famous for his rehabilitation of money as a major determinant of macroeconomic outcomes. For many academic economists, *A Theory of the Consumption Function* (1957) is his greatest work. Friedman showed that the Keynesian concept of household behaviour was fundamentally flawed, arguing that people adjusted their consumption to variations in their long-term expected ('permanent') income. As such, his theory foreshadows the approach to microfoundations that is the cornerstone of modern macroeconomics. His advocacy of economic freedom and market solutions to various socio-economic problems made him a leading policy thinker.

Keywords

Assumptions controversy; Bernoulli, D.; Choice under uncertainty; Consumption function; Econometrics; Economic freedom; Excise taxes; Expected utility hypothesis; Flexible exchange rates; Friedman, M.; Hotelling, H.; Inflation; Keynesianism; Knight, F. H.; Monetarism; Monetary approach to the balance of payments; Monetary transmission mechanism; Money; Money supply; Natural rate of unemployment; Nominal income; Non-accelerating inflation rate of unemployment (NAIRU); Permanent income hypothesis; Phillips curve; Prediction; Price theory; Quantity theory of money; Robbins, L. C.; Sequential sampling; Unemployment; Variance

JEL Classifications

B31

Early Years

Born on 31 July 1912 in New York City, Milton Friedman was the son of a poor immigrant dry-goods merchant, who died when Friedman was 15. Friedman was clearly outside the East Coast establishment of the United States, although he did spend a year in graduate studies at an Ivy League school, Columbia. He graduated (BA) at Rutgers University in 1932 and completed his AM at Chicago in the following year. After a fellowship at Columbia in 1933–4, he returned to Chicago as a research assistant to Henry Schultz to work on demand analysis, until in

1935 he joined the staff of the National Resources Committee. From 1937 he started a long association with the National Bureau of Economic Research (NBER), which continued until 1981. From 1938 he began another long association – with Rose Director, his wife, which produced, *inter alia,* a son and a daughter.

In 1940 there followed a brief period as visiting professor of economics at Wisconsin. Then, after a two-year stint (1941–3) in the Treasury in the division of tax research, he became associate director of the statistical research group in the division of war research at Columbia University, which lasted until the end of the Second World War. He then spent a year as associate professor at the University of Minnesota, before returning to Chicago as professor of economics in 1946, the year in which he received a Ph.D. from Columbia. His teachers at Rutgers were Homer Jones and Arthur Burns; at Chicago, Frank Knight, Lloyd Mints and Jacob Viner; and at Columbia, Harold Hotelling, J.M. Clark and Wesley Mitchell.

Superficially this record does not seem impressive. Yet it encompasses what some scholars, particularly statisticians, would regard as Friedman's most impressive contributions. Inspired by Hotelling's work on the rank correlation coefficient, his first seminal contribution (1937) was the development of the use of rank order statistics to avoid making the assumption of normality in the analysis of variance. After 70 years this article is still regarded as one of the two or three critical papers in the development of nonparametric methods in the analysis of variance, and it was followed by a discussion of the efficiency of tests of significance of ranked data. It is not surprising that these papers have been of considerable practical use, since they were largely a development of Friedman applying his mind to the practical problems he encountered in analysing incomes and consumer expenditure at the NBER and in Washington. Even at this early stage his work bears the imprint that readily identifies all his subsequent work: it is seemingly 'simple', eschewing complexities and complications, concentrating on essentials, and all combined into a lucid exposition.

The detailed analysis of data on incomes and expenditures was Friedman's main occupation during these years. With the exception of Kuznets, Mitchell and Burns, it is difficult to find any eminent economist who acquired such a grounding in the basic empirical material of economics. It is characteristic of all his work that the organization of such data would suggest theoretical developments and new ways of arranging the material, and above all new insights into the economic process. His first published article (1934) was on a method of using the separability of the utility function to measure price elasticities from budgetary data. This exploration of new insights into old data was particularly evident in his book (1945; with Kuznets as joint author) on incomes from private professional practice; there one sees the first signs of the permanent income hypothesis and, indeed, the perceptive reader may guess what is likely to follow. In this book, which Friedman submitted as a doctoral thesis, he argued that the process of state licensure enabled the medical profession more effectively to limit entry into their profession and so enabled them to exploit their patients, keeping fees high and competitors out. The fact that the argument was tightly constructed and buttressed with convincing evidence generated the most vehement opposition and animosity from that proud profession, which appears unabated seven decades later.

Wartime service in the statistical research group, although an interlude in Friedman's basic work on incomes and expenditures, generated one of the most remarkable advances in statistical theory since the seminal contributions of Sir Ronald Fisher. The group was a galaxy, consisting of Abraham Wald, Allen Wallis, Jacob Wolfowitz, Harold Hotelling and many other distinguished statisticians. The sampling inspection of wartime production of munitions and so forth was a tedious process of selecting a sample of a given size and testing to see the fraction of good ones in the batch. Friedman, together with Allen Wallis and Captain Schuyler, observed that testing a given size of sample was clearly wasteful. The process of testing itself gave information that enabled one to determine the degree of confidence achieved. Thus instead of continuing to test up to a fixed size of sample, the testing could be halted whenever a predetermined level of confidence in

the decision had been reached. Friedman formulated the basic idea of what later came to be called 'sequential sampling' and caught the interest and imagination of Wald, who developed and proved the theorem underlying the probability ratio test and eventually produced the influential book Sequential Analysis in 1947. These ideas were adapted very rapidly, and sequential analysis became the standard method of quality control inspection. Like so many of Friedman's contributions, in retrospect it seems remarkably simple and obvious to apply basic economic ideas to quality control; that, however, is a measure of his genius.

At the end of the Second World War, Friedman could have continued his work as a statistician. He would have achieved a stature probably as great as that of his most influential teacher, Harold Hotelling. Alternatively he had all the basic qualifications to take the lead in developing the burgeoning field of econometrics, with its great emphasis on the adaptation of statistical theory to modelling economic phenomena. He chose neither. His excursions into statistics were utilitarian rather than speculative, and he could see little to be gained by the endless sharpening of statistical knives, which was the stuff of econometrics during those years following the Second World War. In this decade, his contributions to statistics were even more intimately linked with his strong belief, implanted largely by Mitchell, that economics could acquire plausibility only by being subjected to empirical verification. In spite of the predilections of many economists, Friedman believed that economics should be viewed as an empirical science.

1946–1955

This decade at Chicago, much influenced by the wisdom of Frank Knight, witnessed the rapid development of economics as a positive science with its own methodology. The prevailing view of economic theory, as developed by Lionel (later Lord) Robbins, was that the veracity of theory could be tested primarily by the correspondence between assumptions and facts. In his 'Methodology of Positive Economics' (in *Essays in Positive Economics,* 1953), Friedman argued per contra that even if one could specify empirical correlates for the assumptions (and this cannot be done in cases where the assumptions are 'ideal types' such as homo economicus), that is irrelevant for judging the usefulness of the theory. Only by the correspondence of the predictions and facts should theories be provisionally accepted or rejected. Results, not assumptions, should be the main focus of our scientific activity in understanding the real world. This approach applied the new philosophy of science, developed by Karl Popper, to economics and by implication to associated social sciences. To countless students, Friedman provided an agenda for what Imre Lakatos later called a progressive research programme. The simplicity of a theory in its ability to explain a lot in exchange for a little input and the degree of 'surprise' in the prediction were the hallmarks of the new approach to theory. But it was in the efficacy and power of the empirical tests that substantial progress was to be made.

In subsequent years the 'Methodology' has been the subject of enormous controversy. There is general agreement that in applying the theory one cannot dismiss the factual basis of the assumptions in quite such a cavalier manner. Furthermore, no one would be rash enough to declare a (refutable) theory discredited if there were a single or a few counter-examples to contradict the predictions. Such absolutism has given way to more subtle interpretations depending, as Lakatos argued, on the new and surprising insights to be obtained. Most theories coexist with small subsets of anomalous results that strictly should discredit them, and yet they remain useful theories and superior to any suggested alternative. But there is no doubt that the substance of Friedman's 'Methodology' has not merely stood the test of time but has also had a profound and lasting effect on the profession.

The application of this methodological approach reached its apotheosis in what most academic economists would regard as Friedman's greatest work, *A Theory of the Consumption Function* (1957). The fundamental proposition that emerged from Keynes's *General Theory* (1936) was that

households expanded their consumption spending by an amount less than the increase in their current income, and that this relationship was sufficiently stable to form the basis for the multiplier through which an increase in autonomous expenditure at the macro level generated a considerably larger increase in real aggregate demand. Since the regularity and predictability of the consumption function was central for the Keynesian control of the economy, it was with trepidation that many observers found that there were considerable inconsistencies between the patterns of household behaviour, particularly from the cross-section data of household surveys and the time series of the historical record. It certainly appeared that the data were quite inconsistent with the Keynesian consumption function. Friedman showed that the Keynesian concept of household behaviour was fundamentally flawed, and that the statistical results suffered from the regression fallacy. People adjusted their consumption with respect to variations in their long-term expected (or 'permanent') income, and paid little heed to transitory variations. This basis idea was not new – indeed it can be found in the 18th-century writings of Bernoulli – but Friedman's development showed his genius for simplicity and for the insights of thinking concretely.

But the main quality of *A Theory of the Consumption Function* was the incomparable amassing, organization and interpretation of the evidence. The relatively low propensities to consume evident in the cross-section data were shown to be entirely consistent with the much higher propensities that emerged from analyses of aggregate time series, when both sets of figures were interpreted in the form of the permanent income hypothesis. Because of the transitory component in the cross-section samples of households, the variance of measured income exceeded the variance of permanent income, and so the slope of the regression of consumer spending on income was much lower than in the aggregate time series regressions, where the transitory component was trivially small. The permanent income hypothesis adequately passed the acid test of using little to explain much.

The integrity of scholarship was demonstrated by the diligent search to find evidence that would discredit the permanent income hypothesis. It was not, and is not, normal practice to scour the literature and statistical evidence for material that might discredit a theory. But Friedman used the hypothesis in the most imaginative way to forecast, for example, the values of regression coefficients for different groups with varying fractions of transitory to permanent income. And he left instructions for other researchers to guide them in tests to be made with further analyses of different data. One of the great contributions of this book was to give a new standard for empirical economics generally. Clearly this was how it *should* be done. The second important effect was the introduction of the concept of permanent income into virtually every field of applied economics, such as monetary economics, housing, transport and international trade. It was a new way of thinking about chance variations and people's decisions in the real world.

A particularly fruitful theoretical approach to the utility analysis of risk and the measurement of utility, based on the work of von Neumann and Morgenstern, appeared in two papers with L. J. Savage (1948, 1952). Using axioms that most observers would regard as acceptable and reasonable, these papers showed that choice under conditions of uncertainty could be represented as a simple process of maximizing expected utility. Thus the utilities of each of the chance outcomes were weighted by the probability of that outcome, and the sum gave an index of expected utility which, given the axioms, would be maximized by choosing from the alternative uncertain prospects. Again the basic idea was not new (it was developed originally by Bernoulli in solving the St Petersburg paradox), but Friedman and Savage discovered new insights and implications, with wide-ranging applications. Apart from rationalizing the widespread practice of simultaneously gambling and insuring, the hypothesis had a profound effect on the theory and practice of portfolio selection. For the pure economic theorist it offered the attractive proposition that, up to an arbitrary linear transformation of origin and scale, utility should be regarded as a cardinal magnitude.

Subsequent discussion (particularly by Maurice Allais) suggested that one of the axioms

(the so-called 'strong independence axiom' which asserted that the preference order would not be affected by mixing these outcomes with equiprobable alternative outcomes) was clearly implausible and violated frequently in practical decisions. Research suggested also that in some fields, for example in air passenger insurance, the expected utility hypothesis was discredited. Nevertheless, the hypothesis still forms a cornerstone of all work – and particularly practical work – in choice among risky alternatives. With some minor exceptions, these papers mark the last contributions of Friedman to the pure theory of statistics and decision-making. Many statisticians regard the diversion of such a fertile mind from its natural field as a great shame and loss.

The gain to empirical economics – and during these years, particularly to the theory of price – was, one suspects, worth the loss. The reformulation of Marshallian demand theory as a practical instrument of analysis (1949) was an exercise in meticulous scholarship in the history of thought, but one which also argued for approaching demand analysis as a positive rather than a normative discipline, an approach which he attributed to Marshall. But the analysis of economic policy, and particularly a critique of the logical structure of the arguments and the empirical evidence adduced to support proposals on economic policy, became increasingly important. Thus the critique of the arguments showing the inferiority of excise taxes compared with alternative income taxes (1952) exposed basic methodological weaknesses in what were the standard treatments of the day.

The demonstration of the uses, as well as some abuses, of the theory of price was one of the highlights of Friedman's lectures of 1946 to 1976 (with a gap from 1963 to 1973), at the graduate school of the University of Chicago. The exploitation of demand and supply as an 'engine of discovery' reached out well beyond those conventionally defined limits of the subject. In these lectures Friedman gave full rein to his persistence and determination to fearlessly pursue the argument, with subtlety and imagination, wherever it led. To the students it opened up new vistas – such as the theory of human capital – and exciting ways of unravelling puzzles and resolving problems. In his hands, economics had both power and point, reality and relevance (for example, 1962). As distinct from much economic work, where complicated ideas are developed in a simple way, Friedman showed how to interpret simple ideas in a most sophisticated way.

This quality characterized his work on money, which, with the inauguration of his monetary workshop in 1951, began to be a major interest for Friedman himself and the distinguished students and faculty that he inspired. The motivations for studying money were firmly implanted when Friedman was at the Treasury dealing with wartime inflation management, but the immediate incentive was the request from the NBER to contribute a study on money for Wesley Mitchell's project on long-term business cycles. Monetary policy as a main tool of macroeconomic management was consistent with a wide degree of free unfettered enterprise and so had an obvious appeal to the liberal (which will be used here in the 19th-century sense) Friedman. The prevailing Keynesian orthodoxy, with its emphasis on expanding the public sector, appeared to threaten liberal society. The Post Keynesian contempt for money was a tempting target that was difficult to resist. But undoubtedly Friedman's imagination had been challenged by the Chicago School's preference (particularly by Knight and Simons) for rules rather than authorities in macroeconomic as well as microeconomic policy. The uncertainties of the economic environment would be much reduced if the Federal Reserve Board followed simple rules. Friedman first suggested (1948) a countercyclical rule of financing recession-induced increases in the federal budget deficit by money creation and correspondingly by retiring money during a boom-induced surplus. The empirical evidence that he explored in subsequent years, however, led him to formulate the rule of a fixed and known expansion of the money stock, rather than indulging in countercyclical operations in vain attempts to stabilize the economy. Whatever his motives, however (and one should note that motives are quite irrelevant in judging substantive propositions), for the next 30 years Friedman's work was focused on money. At last monetary

economics was to be interpreted as part of the central corpus of price theory; it was to be integrated into economics.

The Monetary Revolution and the Rise of Monetarism, 1956–1975

In the late 1950s, to anyone subjected to the Anglo-Saxon schools of economics during the previous two decades any attempt to revive monetary economics appeared to be foolhardy, like flogging a decomposing horse. The Radcliffe Committee, advised by the most eminent economists, had reported in 1959 that the quantity of money was of little or no interest since the velocity of circulation had no limits. The quantity theory of money was subject to particular scorn as a mere identity without content. As Friedman was to point out, however, all theory consists of tautologies; all that theory does is to rearrange the implications of the axioms to produce interesting, even surprising, consequences. But they remain empty and devoid of substantive as distinct from speculative content, until they have been tested against a wide body of facts.

Of course, for many years the quantity theory of money had been tested against experience and data and over several critical periods of change. The most distinguished exponents of such tests had included Irving Fisher and Keynes himself, as well as the irrepressible Clark Warburton. Yet the methodology was murky, the statistics slim, and interrelationships between data and theory obscure. In *Studies in the Quantity Theory of Money* (1956), Friedman and his co-authors redefined the quantity theory in terms of statements specifying a degree of stability in the demand for money. It was proposed that the demand for money by the individual household would be a stable function of its money income (later thought to be permanent income or wealth) and the cost of holding money represented by the rate of interest and the expected rate of inflation.

Friedman's presentation of the theory of the demand for money in the first essay in *Studies* is one of his most widely quoted papers, primarily because it is thought to show that in presenting the money demand function as a portfolio decision with respect to alternative assets, rather than a demand related to the flow of transactions and income, Friedman was a closet Keynesian. Substantively this was a side issue; the main point was the stability of demand, particularly with respect to nominal income or wealth. Unfortunately, this first essay was not one of Friedman's better expositions. The other essays in *Studies,* particularly that of Cagan on hyperinflations and Selden on velocity, however, established the value of examining nominal income and inflation in the context of the demand for money. The quantity theory in its new reborn Chicago form had passed its first tests.

The unknowns, however, remained legion. The vexed question of the nature of the regime controlling the supply of money, and how to interpret the problem of identifying the demand function in the data were to persist, in the eyes of many critics, as the major weakness in such studies. Was the stock of money reacting passively to changes in nominal income (or wealth) or were prices and output responding to endogenous changes in the supply of money? The Chicago workshop averred that the answer to such questions could be obtained only by painstaking research into the history of the monetary process. Undoubtedly there were occasions when the money stock passively responded to changes in nominal income, but equally obvious were instances where the money supply changed for reasons quite independent of past or contemporaneous movements in money incomes. The role of the balance of payments and the exchange rate regime was clearly recognized, and it is not difficult to discover the genesis of the monetary theory of the balance of payments in 'Real and Pseudo Gold Standards' (1961) and other essays in *Dollars and Deficits* (1968).

Although the detailed development of the history of the money supply process and the relationships with gold and exchange rates were to appear in the monumental *A Monetary History of the United States, 1867–1960* (1963), Friedman had already made it perfectly clear that a stable growth of the money supply was unlikely to be feasible under a regime of fixed exchange rates.

His advocacy of flexible exchange rates (in 1953) followed logically on his views of the efficacy of free markets. Friedman was one of the very few economists (Gottfried Haberler and Egon Sohmen were among them) who clearly showed that the ambient dollar shortage was merely a consequence of fixed exchange rates and divergent monetary policies. His analysis was amply justified when by the 1960s, due to the change in monetary policies, the dollar shortage had turned into a dollar glut.

Yet in spite of the increasing attention paid to the balance of payments and the money supply process generally, the prime focus of Friedman's work remained the examination of the effects of monetary variations on nominal income, prices and output. The main questions were: (*a*) what was the relative importance of monetary compared with fiscal variations (the Keynes versus Monetarist debate); (*b*) what was the time pattern of adjustment; and (*c*) could expansionary financial or fiscal policies affect real output in the short or long run? The answers that evolved from Friedman's analysis were: to (*a*), although an increased fiscal deficit had an impact effect on nominal income this soon disappeared, whereas after a lag the increased rate of money growth permanently augmented the rate of inflation; to *(b),* the adjustment of nominal income to an increased rate of monetary growth involves lags that are 'long and variable'; and to (*c*), in the long run additional monetary growth affects only the rate of inflation and has virtually no effect on either the level of output or its growth rate. In essence Friedman found that variations in the rate of growth of the money supply had short-run effects – sometimes, as in 1931 of a devastating magnitude – on real output as well as on prices; but in the long run (more than three years) the only substantial effect was on prices.

Over the 1960s and 1970s the results of Friedman's research for the long run were widely accepted. The logic as well as the data were appealing: nominal variations (in money) have nominal effects (on prices) and no real effects (on output). But such agreement did not readily extend to his short-run claims of, first, the impotence of fiscal policy in countering cyclical oscillations and shocks; and, secondly, the large but unpredictable effects of monetary variation on real output and employment. The claims of Keynesian economists for the stability and size of the fiscal multipliers continued, but it is noteworthy that estimates of the size of the multipliers, except for those produced by the Cambridge (England) School, were substantially reduced in the 1980s. (One is not able to determine whether the economists or the economies have become less Keynesian and more monetarist.)

One of the abiding criticisms of Friedman's work on money (much of it in joint authorship with Anna Schwartz) is that it has no theoretical structure – or more charitably that such theoretical structure as exists is implicit rather than explicit. Processes of monetary transmissions as he describes them are alleged to be 'black boxes' with no precise specification of the way in which money works its magic. Friedman attempted to produce a theoretical underpinning for his approach to research in *Milton Friedman's Monetary Framework* (1974) by producing a seven-equation basic model of the (closed) economy. The critical difference between the Keynesian and the classical models was the choice of the last equation; the Keynesians chose to specify the price level as fixed by exogenous forces and the level of output as a variable determined by the level of aggregate demand, whereas the classical economists held that the level of real output was fixed by technology, skill and so on, and that the price level was determined by the model. With this simple model, Friedman was able to highlight the differences of method and approach as primarily different views about the size and stability of the coefficients of the system. In principle, at least, such issues could be resolved by appeals to the evidence. *The Framework* did not, however, make substantial progress in providing a sound analytical basis for the dynamics of the adjustment, through output, price and interest rate effects, to the new long-run equilibrium. The transmission mechanism and dynamics remain enshrouded in the gloom of a black box.

Yet in spite of what many theoretical economists considered to be drastic limitations for sound theoretical developments, in the most

important and influential paper in macroeconomics in the post-war years, his presidential address to the American Economic Association, Friedman showed that the view of macroeconomic policy as a trade-off between unemployment and inflation was fundamentally flawed (1968). In the long run there was no such trade-off, while in the short run the tradeoff took place only during the adjustment to the new inflationary environment, and then only because people were temporarily surprised by the new environment. The overriding objective of contractual arrangements was to fix real wages and prices. Money served as a veil, sometimes seductive but always obscuring underlying reality. The so-called Phillips curve was a short-term temptation rather than a long-term choice.

Friedman caught opinion at ebb and turned it into a flood. Throughout the 1960s the trade-off between unemployment and inflation appeared more and more illusory. Unemployment went up but inflation did not go down; it also increased. Into the 1970s and particularly during the great inflationary recession of 1974/75, when both inflation and unemployment reached new highs in most Organisation for Economic Co-operation and Development (OECD) countries, it appeared that only Friedman's view made any sense. Like Keynes's *General Theory,* it was one of the very few contributions that changed both the approach of professional economists and the policies adopted by finance ministers. Some time during the 1970s most governments recognized that the road to fuller employment did not lie over the high sierra of soaring inflation. Doctrinally, economists took into their toolbox the Friedman concept of a 'natural rate' of unemployment where inflation would neither accelerate nor decelerate. (The word 'natural', which was usually considered either normative or even desirable, was generally eschewed in favour of the term 'non-accelerating inflation rate of unemployment' or NAIRU.)

The natural level of unemployment was held to be determined by the nature of labour markets, such as the conventions of wage contracts, the degree of mobility, the level of unemployment benefits, the marginal utility of income, and many other 'structural' factors that are independent of the rate of inflation. As in the case of the permanent income hypothesis, to which it is distantly related, the concept had applications in fields far from labour markets. At the same time it provided one of the many missing links between the macroeconomics of aggregate output and inflation and the microeconomics of industrial adjustment and resource allocation. Again, in retrospect it all seems obvious; but that merely measures the magnitude of the contribution.

By any standards – even those of Keynes and the *General Theory* – Friedman's contribution to monetary analysis and policy must be ranked very high. Every economist, finance minister and banker felt his influence. But, as an accomplishment of the intellect, one suspects that most of Friedman's peers would still regard his work on the consumption function as the maximum maximorum of his contributions to economics. Friedman's monetary analysis did not have that sense of comprehensiveness and structural balance that are the hallmarks of his work on consumer spending. One closed *A Theory of the Consumption Function,* not with the feeling that nothing more need be said, but that whatever was discovered in the future must fit neatly into this superb and satisfying framework. The architecture could accommodate, and indeed so far has shaped and absorbed, all new contributions. The *Monetary History* and the *Framework,* however, although probably more influential in doctrine and policy, did not provide the commodious and harmonic form of the Consumption Function. A number of awkward corners left one wondering what to do. And since the theoretical plans were left obscure, sometimes there were questions whether the superstructure would really hold up. But this does not belittle the *Monetary History* so much as praise the *Consumption Function.*

1975–2006

The award of the Nobel Prize for Economics, long overdue in 1977, at last recorded that Friedman's great contributions had even penetrated the Swedish academies. Inevitably Friedman's rise to stardom had given many more opportunities to persuade electorates through the medium of the

popular press (highlights include his columns in *Newsweek* from 1966 to 1984) and television (in the popular PBS and BBC series *Free to Choose* in 1980). His contributions to persuasive journalism delighted many, infuriated some, and made all his serious readers, if not wiser, then certainly better informed. In all these popular articles the high professional standards of integrity were maintained. But at the same time Friedman continued with his scholarly work on monetary analysis; examples include Friedman (1988, 1990, 1992). *Perhaps the main output, after more than 20 years of effort, was his book with Anna J. Schwartz,* Monetary Trends in the United States and the United Kingdom, Their Relation to Income, Prices and Interest Rates, 1867–1975 (1982).

The main methodological decision lying behind this study was that, since there was too much inexplicable variation in short-run variations in income and money, it was best to ignore these and concentrate on comparing the cyclical phase averages. These would screen out the short-term effect and would enable an analysis to be made of the underlying long-term money–income–interest relationships. Even for this team of Friedman and Schwartz, the treatment of the data and the integrity of their analysis reached new heights of meticulous scholarship.

Yet, considering the enormous value of the input of time and energy, the results are, as the authors confess, hardly worth the cost. For the most part the study confirms, and demonstrates with comparative data for the United States and the United Kingdom, the basic propositions on velocity, real income, prices and interest rates that had emerged in the *History.*

In his final decades, it may be claimed that Friedman had fallen prey to the same temptations that affected Alfred Marshall. For many years of his mature professional career, Marshall spent much of his time revising and refining his great *Principles.* In retrospect it seemed to be a great loss to scholarship that Marshall did not leave the *Principles* well alone and turn to his projected study of the economics of the state. The opportunity was missed. It would be, however, a travesty to draw a close parallel between Friedman and Marshall in their mature years. Perhaps with the example of Marshall in mind, Friedman had generally launched his studies on the profession and then left them largely to fend for themselves. (The only exception is the textbook *Price Theory: A Provisional Text,* 1962, which was revised in 1976.) Yet there is a sense in which Friedman, trapped by his immense success in monetary economics, had been prevented from deploying his mind in scholarly work in other fields of economics.

The possibilities are revealed in Friedman's more popular writings on issues such as public spending, price and rent control, taxation, and many issues in microeconomics. Characteristic flashes of insight and phrase, together with the innovations of approach – especially the simplifications – give the professional reader a tantalizing taste of what might have been yet another great contribution to economic science. Many economists have always believed that, in spite of his great strides in money, Friedman's relative advantage was always in the study of price theory and its manifest applications. There is the measure of the man.

The Public Image of Friedman

The conventional view of Friedman is that he is one of the most ardent and most effective advocates of free enterprise and monetarist policies over the six decades 1945 to 2006. If far short of his wishes, the success of his advocacy has by any objective standard been enormous. Opinion in Western countries, even among the clerisy, has moved decisively in its preference for those economic freedoms that he has so eloquently advocated.

It is not possible to parcel out any neat attribution of influence on these great changes in attitude and policy. Friedman himself would probably give by far the largest weight to the experience of the 1970s, particularly the disappointments over failure to restrain the growth of government spending and the great inflation from 1965 to 1981. The explanation of these events and the development of an alternative strategy, with institutions that would ensure individual economic liberty and freedom from inflation, have been, in the public perception, Friedman's great

contribution to the reforms. In his appearances in the various media he was a great persuader, and he played a critical role in promoting such ideas as an all volunteer army, the voucher schemes for education and health, and indexing income tax. In effectiveness, breadth and scope, his only rival among the economists of the 20th century is Keynes.

See Also

- ▶ Monetarism
- ▶ Quantity Theory of Money

Selected Works

1934. Professor Pigou's method for measuring elasticities of demand from budgetary data. *Quarterly Journal of Economics* 1: 151–163.

1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32: 675–701.

1940. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics* 11: 86–92.

1945. (With S. Kuznets.) *Income from independent professional practice*. New York: NBER.

1948a. (With L.J. Savage.) The utility analysis of choices involving risk. *Journal of Political Economy* 56: 279–304.

1948b. (With H.A. Freeman, F. Mosteller, and W. Allen Wallis.) *Sampling inspection.* New York: McGraw-Hill.

1949. The Marshallian demand curve. *Journal of Political Economy* 57: 463–495. Repr. in (1953).

1952. (With L.J. Savage.) The expected utility hypothesis and the measurability of utility. *Journal of Political Economy* 60: 463–474.

1953. *Essays in positive economics.* Chicago: University of Chicago Press.

1956. *Studies in the quantity theory of money.* Chicago: University of Chicago Press.

1957a. (With G.S. Becker.) A statistical illusion in judging Keynesian models. *Journal of Political Economy* 65: 64–75.

1957b. *A theory of the consumption function.* Princeton: Princeton University Press.

1959. *A program for monetary stability.* New York: Fordham University Press.

1962a. *Capitalism and freedom.* Chicago: University of Chicago Press.

1962b. Price theory: A provisional text. Chicago: Aldine.

1963a. (With A.J. Schwartz.) *A monetary history of the United States, 1867–1960.* Princeton: Princeton University Press for the NBER.

1963b. (With D. Meiselman.) The relative stability of monetary velocity and the investment multiplier in the United States, 1897–1958. In *Stabilization policies.* Englewood Cliffs, NJ: Prentice-Hall.

1967. (With R.V. Roosa.) *The balance of payments: Free versus fixed exchange rates.* AEI Rational Debate Seminar. Washington, DC: American Enterprise Institute.

1968a. The role of monetary policy. Presidential Address, American Economic Association, 29 December 1967. *American Economic Review* 58: 1–17. Repr. in (1969).

1968b. Dollars and deficits: Inflation, monetary policy and the balance of payments. Englewood Cliffs, NJ: Prentice-Hall.

1969. The optimum quantity of money and other essays. Chicago: Aldine.

1970. (With A.J. Schwartz.) *Monetary statistics of the United States.* New York: Columbia University Press for the NBER.

1972a. (With Wilbur J. Cohen.) *Social security: Universal or selective?* AEI Rational Debate Seminar. Washington, DC: American Enterprise Institute.

1972b. An economist's protest: Columns on political economy. Glen Ridge, NJ: Thomas Horton.

1973. *Money and economic development.* Horowitz Lectures of 1972. New York: Praeger.

1974. Milton Friedman's monetary framework: A debate with his critics, ed. R.J. Gordon. Chicago: University of Chicago Press.

1976. *Price theory.* Chicago: Aldine. (Revised and enlarged version of the 1962 edn.)

1978. Tax limitation, inflation and the role of government. Dallas: Fisher Institute.

1980. (With R. Friedman.) *Free to choose.* New York: Harcourt Brace Jovanovich.
1982. (With A.J. Schwartz.) *Monetary trends in the United States and the United Kingdom.* Chicago: University of Chicago Press, for the NBER.
1984. (With R. Friedman) *Tyranny of the status quo.* San Diego, New York and London: Harcourt Brace Jovanovich.
1988. Money and the stock market. *Journal of Political Economy* 96: 221–245.
1990. Bimetallism revisited. *Journal of Economic Perspectives* 4(4): 85–104.
1991. (With A. Schwartz.) Alternative approaches to analyzing economic data. *American Economic Review* 81: 39–49.
1992. Do old fallacies ever die? *Journal of Economic Literature* 30: 2129–2132.
1992. *Money mischief: Episodes in monetary history.* San Diego and London: Harcourt Brace Jovanovich.
1998. (With R. Friedman.) *Two lucky people: Memoirs.* Chicago: University of Chicago Press.

Bibliography

Hammond, J.D. 1999. *The legacy of Milton Friedman as teacher*. Cheltenham: Edward Elgar.
Hammond, J.D. 2005. *Theory and measurement: Causality issues in Milton Friedman's monetary economics*. Cambridge: Cambridge University Press.
Keynes, J.M. 1936. *The general theory of employment, interest, and money.* New York: Macmillan.
Wald, A. 1947. *Sequential analysis*. New York: John Wiley.
Wood, J. 1990. *Milton Friedman; Critical assessments*. London: Routledge.

Friend, Irwin (1915–1987)

Paul Taubman

Friend was born in Schenectady, New York and received his PhD from American University in 1953. He then became a professor of finance and economics in the Wharton School, University of Pennsylvania, being president of the American Finance Association in 1972.

His many books and articles deal with securities markets, financial institutions, tax policy, capital asset pricing, consumption and saving functions, econometric models, and the usefulness of expectations and anticipations data. Perhaps Friend's most important contribution to economic theory per se is in 'The Demand for Risky Assets', (1975), the first part of which greatly extends the capital asset pricing model in several directions. These include the explanation of the determinants of the basic risk premium between risky assets as a whole and the riskfree rate, incorporating income taxes, allowing the riskfree asset to have a positive supply, and allowing for human capital. The paper demonstrates how the micro theory can be aggregated to obtain a macro model suitable for testing with the time series data, as is done in the latter part of the article.

It is generally recognized that theorists assume away many problems in order to highlight a central issue. The restrictions imposed by theorists often provide strong conclusions, which would not hold without the restrictions. Friend has often tested these restrictions and his results require changes in theory. For example, in the paper just cited, the question is posed of whether typical investors have increasing or proportional risk aversion. The paper presents fairly strong evidence that the appropriate utility function should have proportional risk aversion as part of its properties and provides measures of risk aversion for the market as a whole. Similarly, Friend pushed the permanent income hypothesis to its limits and helped refine it in 'Consumption Patterns and Permanent Income', showing it is not valid to restrict the marginal propensity to consume out of transitory income to be zero. Other areas in which his work has seriously questioned the usefulness of basic assumptions almost universally made by theorists include the supposed irrelevance of unique risks in the pricing of risky assets, the applicability of the customary factor analysis in confirming the arbitrage pricing theory, and the complete faith of many economists in

the stock market's efficiency and the undesirability of any form of government intervention (including mandated disclosure).

Selected Works

1957. (With I. Kravis). Consumption patterns and permanent income. *American Economic Review* 47(2): 536–555.
1964. (With P. Taubman). A short-term forecasting model. *Review of Economics and Statistics* 46(3): 229–236.
1975. (With M.E. Blume). The demand for risky assets. *American Economic Review* 65(5): 900–922.
1981. (With R. Westerfield). Risk and capital asset prices. *Journal of Banking and Finance* 5(3): 291–315.
1983. (With J. Hasbrouck). Saving and after-tax rates of return. *Review of Economics and Statistics* 65(4): 537–543.
1984. Economic and equity aspects of securities regulation. In *Management under government intervention: The view from Mt. Scopus*, ed. Lanzillotti and Peles. Greenwich: JAI Press.
1985. (With P.J. Dhrymes, M.N. Gultekin and N.B. Gultekin). New tests of the APT and their implications. *Journal of Finance* 40(3): 659–674.

Frisch, Ragnar Anton Kittel (1895–1973)

P. Nørregaard Rasmussen

Keywords

Acceleration principle; Clark J. M.; Demand analysis; Econometric society; Econometrics; Frisch R. A. K.; Input–output analysis; Institute of Economics (Norway); Macroeconomics, theory of; Marshall, A.; Methodology of economics; National accounting; Production, theory of; Statistics and economics

JEL Classifications

B31

Frisch lived a long, varied and extremely productive life. He graduated in economics at the University of Oslo in 1919 (although as a son of a goldsmith he 'supplemented' this by finalizing his apprenticeship as a goldsmith in 1920!). He studied in France from 1921 to 1923 and in Britain in 1923; was an associate at the University of Oslo from 1925 and received his doctorate in 1926 in mathematical statistics (Frisch 1926a). Further studies abroad in the USA, France and Italy (1927–8) were followed by an associate professorship at the University of Oslo (1928) and a full professorship in 1931. Frisch was head of the (newly established) Institute of Economics in Oslo from 1932 to his retirement in 1965. He was also chief editor of *Econometrica* (1933–55), followed by his chairmanship of the editorial board. He was one of the founders (1930) and, in fact, the driving force behind the creation of the Econometric Society. He was a member of a number of national and international expert committees and adviser on several occasions to developing countries (India 1954–5 and Egypt several times over the years 1957–64). He received honorary doctorates from a number of universities (*inter alia* Stockholm, Copenhagen, Cambridge, Birmingham) and was – together with Jan Tinbergen – the first (1969) to receive the Alfred Nobel Memorial Prize in Economics. In addition he received (as the first recipient) the Schumpeter Prize (1955), the Feltrinelli Prize (1956) and three *Festschriften*. He was a visiting professor or guest lecturer to a number of universities – Yale, Minnesota, Paris, Pittsburgh, for example – and he was a very active participant at numerous international meetings of economists, statisticians and mathematicians. In the late 1940s there was a joke among Norwegian students that he was also a 'visiting' professor in Oslo. This was unfair. In particular during the 1930s he put a lot of effort into his teaching and

was writing a series of lecture notes, most of them seminal, though many remained unpublished. The impressive list of his publications (Haavelmo 1973) and activities could be continued because he was a genius, cutting through problems like a warm knife through butter, and because his working power was extraordinary.

To survey his contributions is not easy for the simple reason that there is scarcely any area of economics Frisch has not been into and left his imprint. To cooperate with him was not always easy. He was too strong, as shown by the fact that he seldom had co-authors. The list of his published and printed works comprises about 160 items. But to this should be added a long series of mimeographed contributions – many will recall the 'Memoranda fra Social konomisk Institutt' – from about 1946 onwards. They amounted altogether to 6,500 pages, and most of them are still awaiting publication (though Frisch himself argued that 'for editing it needs a very good man, and if he is good enough, he should write himself').

Frisch began his academic work in the theory of mathematical statistics. This profession today acknowledges his early contributions and regrets his departure from it, though admitting that in terms of the more applied theory of statistics he made noticeable contributions later on. His years in Paris, where he concentrated on mathematics, were not in vain.

It is, however, in economics that Frisch made his name, He was at most of the centres and many of the corners of the subject. One may, however, also argue that his most significant contribution is in economic methodology. This comes out not only in his applications of methods but also in their general presentation. A very good example was written overnight in a hotel room at Colmar, after a day's discussions at a meeting of the Econometric Society. The article (Frisch 1936b) is a classic, clearing the ground about the very meaning of static versus dynamic analysis. This is by now elementary, but it is elementary because of Frisch. In his principal works on methodology, he used and unified the tools he had mastered so well: economic theory, mathematics and statistics. It is no accident that he invented the word 'econometrics', for in general he enriched our methodological vocabulary by a number of precise concepts: macro- versus micro-analysis, statics versus dynamics, exogenous versus endogenous variables, the concept of autonomous relations, the problem of identification of relations, confluent relations, decision models, conjectural behaviour (of firms) – a complete list would be very long.

Few would hesitate to agree that Frisch 'created' econometrics in the modern sense of the word. It is much more notable that he warned again and again against misuses of the new tools. In the first issue of *Econometrica* in 1933 he wrote: 'The policy of *Econometrica* will be as heartily to denounce futile playing with mathematical symbols in economics as to encourage their constructive use.' In Frisch (1970), he argued that 'the econometric army has now grown to such proportions that it cannot be beaten by the silly arguments that were used against us previously. This imposes on us a *social and scientific responsibility* of high order in the world of today' (p. 153). But in the very same article he also stressed (p. 163) that 'I have insisted that econometrics must have relevance to concrete realities – otherwise it degenerates into something which is not worthy of the name econometrics, but ought rather to be called playometrics'.

Always underlying Frisch's contributions to methodology were his consistent efforts to turn economics into a precise science, quantifying the variables and the structures. This is different from the traditional 'on the one hand and on the other', where on balance the answer is left in the air. But this 'aggressive' view also presents new challenges. The economist must be prepared for the troublesome work of gathering data, to face the difficulties in estimating structures and in the end to attempt a balanced interpretation of the outcome. Frisch saw this and contributed to this debate throughout his career; illustrations might be Frisch (1933a, 1934b, 1936a, 1939). These and many other contributions had a profound influence and wide applications in pre-war as well as post-war econometrics. Again, and sadly enough, one could also refer to a number of unpublished papers, though these were influential as contributions to scientific gatherings. A supreme example

is a paper on 'Statistical versus Theoretical Relations in Macrodynamics', a contribution to a conference sponsored by the League of Nations in 1938 to discuss Jan Tinbergen's work for the League of Nations on the trade cycle.

One may wonder why Frisch left so many path-breaking contributions without taking the trouble to publish them. I think there is a double answer. On the one hand Frisch was an impatient man: if he had given the gist of the solution to a problem, he tended to go on to new problems. On the other hand, he was extremely careful: a publication going to the printer had to be perfect and finalized – a troublesome process which he often tended to avoid. Haavelmo (1973), reports that Frisch often argued that proofreading is one of the most difficult and important tasks of a scientist.

The general assessment above can be verified by considering Frisch's contributions in the field of demand analysis, the theory of production and the theory of macroeconomics.

In demand analysis he began as early as 1926 (Frisch 1926b), by formulating a number of basic axioms and from these to deduce a theory of demand. Thus utility functions were not postulated but were derived from more basic axioms, all of these being formulated as being, in principle, open to testing. It may be fair to say that his work in this field culminated in Frisch (1959). It is a tribute to his work that it has in fact been used in practice, for example in Norwegian planning.

In the theory of production Frisch was a forerunner, formulating the theory in a strict mathematical form but also applying it on concrete problems. An example is Frisch (1935). However, most of his works were in the form of mimeographed lecture notes in the 1930s and remained unpublished until 1962 and later (Frisch 1962, 1963). The main results, however, were internationally known through the works of Schneider and Carlson, who at times were research associates in Oslo and very much influenced by Frisch.

Also in the theory of macroeconomics, Frisch was at the front, even, it can be argued, ahead of Keynes. Anybody reading his booklet, Frisch (1933b), and the subsequent articles in *Econometrica* (1934a), might be willing to argue that Frisch made it first. He shows convincingly how a capitalist economy may go into a deadlock when, to put it in a simple way, the tailor cannot sell to the shoemaker because the shoemaker cannot sell to the tailor:

> . . . the cause of great depressions, such as the one we are actually in, is . . . fundamentally connected with the fact that modern economic life has been divided into a number of regions or groups.
>
> Under the present system, the blind 'economic laws' will under certain circumstances, create a situation where these groups are forced mutually to undermine each other's position. Each group is forced to curtail the use goods produced and services rendered by the other groups, which, in turn, will cause a still further contraction of the demand for its own products, and so on. (Frisch 1934a, pp. 259f.)

He also, in the 1934 articles, outlined (a couple of years before Leontief) an input–output analysis. His contributions in these areas were not appreciated at the time, but from a historical perspective they are path-breaking. This also holds for his contribution to the (famous) Cassel *Festschrift* (Frisch 1933c), where a dynamic system for the economy as a whole was outlined and where he distinguished in a sharp and fruitful way between the impulses and the propagation mechanism. In this context one may also, as an illustration of his interest in the development of economic theory, make a reference to his excellent analysis of Marshall (Frisch 1950).

There is a direct line from here to his systems of national accounts (first published in Frisch 1939) which had a profound influence on the planning in Norway and elsewhere after the Liberation. In the context of macroeconomics it is illustrative to mention Frisch's discussion with J.M. Clark over the acceleration principle (Frisch 1931, 1932). In an amazingly simple way Frisch cleared up the issue, that is, the interplay between the pure acceleration principle and the reinvestment cycle, as the following quotation shows:

> Let z be consumer-taking [in present day language this is simply consumption] per unit of time, w capital production per unit of time, and W the capital stock that exists at any moment of time. All the three magnitudes z, w, and W are, of course, functions of time. In practice they would be represented by time series.

> Let us, for simplicity, make the following two assumptions: A. Consumer-taking z is the same as the production of the consumer good, and this again is at any time proportional to the existing capital stock W. In other words, we have

$$W = kz, \tag{1}$$

where k is a constant independent of time. B. The depreciation per unit of time u, that is to say, the capital production that is needed for replacement purposes, is proportional to the existing capital stock. In other words, we have

$$u = hW, \tag{2}$$

where h is a constant independent of time. Now, the rate of change with respect to time of the capital stock is equal to

$$\dot{W} = w - u \tag{3}$$

By virtue of (1) we have, however,

$$\dot{W} = k\dot{z}, \tag{4}$$

where $\dot{z}$ is the rate of change of consumer-taking. Inserting this into (3), and expressing u in terms of z by (2) and (1), we get

$$k\dot{z} = w - khz.$$

So that we finally have

$$w = k(hz + \dot{z}). \tag{5}$$

The rate of change with respect to time of capital production is thus equal to

$$\dot{W} = k(h\dot{z} + \ddot{z}). \tag{6}$$

Formula (5) indicates the two parts of which total capital production is made up. In the first place we have the part khz that represents capital production for replacement purposes. This part is (under our simplified assumption) proportional to the *size* of consumer-taking. In the second place, we have the part $k\dot{z}$ representing capital production for expansion purposes. This part is (under the present simplified assumption) proportional to the *rate of change* of consumer-taking. Thus there are two forces that act upon total capital production. If consumer-taking is increasing, but at a constantly decreasing rate, the first of these two forces tends to increase, and the second tends to slow down capital production. Which one of the two forces shall have the upper hand depends on the *manner* in which the increase in consumer-taking slows down, and it depends also on *the rate of depreciation* (Frisch 1931, pp.647 f.).

As will be seen, it is all so simple, *provided* the problem is *formulated* clearly. And formulating problems in a fruitful way was one of his secrets.

What is a genius? It might be argued that Frisch up till now was one of the ten in our profession in the 20th century. Not that he cannot be criticized. On occasion he used his brains more or less in vain, for example, on unimportant calculating schemes. Long after electronic calculators were on the market, he used time and effort on inventing schemes for inverting a matrix on a simple desk calculator. His various methods for the solution of programming problems – 'the logarithmic potential method', 'the multiplex method' and 'the nonplex method' (for example, Frisch 1956, 1957, 1961a, b) – are still disputable, taking present-day techniques into account. In other words, he might have had a weak point in not always being able to evaluate the importance of a problem; that is, he might now and then have used his immense working power on issues where his opportunity costs were too high.

Even so, his life work is impressive. And so was the man himself. His political attitude was rather to the left than to the right – while at the same time he was a devout Christian. He felt a strong social responsibility, as proved through his work on the problems of the 1930s as well as, and perhaps even more so, by his consciousness towards the less developed countries. He could at times be a bit harsh on colleagues who did not live up to his own standards for serious work. At the same time he was extremely kind and helpful to students doing their best. He never failed to encourage. And few will forget when his strong blue eyes were shining with joy.

Selected Works

1926a. Sur les semi-invariants et moments employésdansI'étude des distributions statistiques. *Skrifterutgittavdet Norske Videnskaps-Akademi i Oslo II.Hist.-Filos. Klasse 1926*, No. 3.

1926b. Sur unproblèmed'économie pure. (An attempt at developing an axiomatic foundation of utility as a quantitative notion and at measuring statistically the variation in the marginal utility of money on the basis of data from the Paris Cooperative Society.) *Norsk, matematisk forenings skrifter* Series I 16: 1–40.

1931, 1932.The Interrelation between capital production and consumer taking. *Journal of Political Economy* 39(5) (1931), 646–654. A rejoinder, 40(2) (1932), 253–255.A final word, 40(5) (1932), 694.

1933a. *Pitfalls in the statistical construction of demand and supply curves*. Leipzig: H. Buske.

1933b. *Sparing og Cirkulasjonsregulering*. Oslo.

1933c. Propagation problems and impulse problems in dynamic economics. In *Economic essays in honor of Gustav Cassel*. London: Allen &Unwin. Reprinted in *Readings in business cycles*, ed. R.A. Gordon and L.R. Klein. London: Allen &Unwin, 1966.

1934a. Circulation planning. *Econometrica*, Pt I 2(3): 258–336; Pt II, 2(4): 422–435.

1934b. *Statistical confluence analysis by means of complete regression systems*. Oslo: University Institute of Economics.

1935. The principle of substitution. An example of its application in the chocolate industry. *Nordisk Tidsskrift for Teknisk Økonomi* 12–27.

1936a. Annual survey of general economic theory: The problem of index numbers. *Econometrica* 4(1): 1–38.

1936b. On the notion of equilibrium and disequilibrium. *Review of Economic Studies.* 3(2): 100–105.

1939. Nasjonalregnskapet. *Transactions of the Nordisk StatistikerMøte*, Oslo. 1943. *Økosirksystement. 1943. Ekonomisk Tidskrift* 45: 106–121.

1950. Alfred Marshall's theory of value. *Quarterly Journal of Economics* 64: 495–524.

1952. Frisch on Wicksell. In *The development of economic thought*, ed. H.W. Spiegel. New York: Wiley.

1956. La résolution des problèmes de programmelinéairepar la méthode du potential logarithmique. In *Séminaireéconométrique*, ed. R. Roy. Paris: CNRS.

1957. The multiplex method for linear programming. *Sankhya* 18(3–4): 329–360.

1959. A complete scheme for computing all direct and cross demand elasticities in a model with many sectors. *Econometrica* 27(2): 177–196.

1961a. Mixed linear and quadratic programming by the multiplex method. In *Money, growth and methodology, Fesiskrifttil Johan Akerman*, ed. H. Hegeland. Lund: Gleerup.

1961b. Quadratic programming by the multiplex method in the general cases where the quadratic form may be singular. *Bulletin of the international statistical institute* 38(4) (Tokyo): 283–332.

1962. *Innledning til Produksjonsteorien*. Oslo: Universitets Forlaget.

1963. *Lois techniques et économiques de la production*. Paris: Dunod. Trans. as *Theory of production*. Dordrecht: D. Reidel; Chicago: Rand McNally, 1965.

1970. Econometrics in the world of today. In *Induction, growth and trade; essays in honour of Sir Roy Harrod*, ed. W.A. Eltis, M.F. Scott, and J.N. Wolfe. London: Clarendon Press.

Bibliography

Arrow, K.J. 1960. The work of Ragnar Frisch, econometrician. *Econometrica* 28: 175–192.

Edvardsen, K.N. 1970. A survey of Ragnar Frisch's contributions to the science of economics. *The Economist* 118 (2): 174–196.

Edvardsen, K.N. 2001. *Ragnar Frisch: An annotated bibliography.* Report 4/2001. Ragnar Frisch Centre for Economic Research, University of Oslo. Online. Available at http://www.frisch.uio.no/pdf/rapp01_04.pdf. Accessed 9 Nov 2006.

Haavelmo, T. 1973. Minnetale over Professor, dr. philos. Ragnar Frisch. *Arbokdel Norske Videnskabs-Akademi i Oslo*.

Johansen, L. 1969. Ragnar Frisch's contributions to economics. *The Swedish Journal of Economics* 71 (4): 302–324.

F

Full and Limited Information Methods

Thomas J. Rothenberg

JEL Classifications
C3

Econometricians have developed a number of alternative methods for estimating parameters and testing hypotheses in simultaneous equations models. Some of these are limited information methods that can be applied one equation at a time and require only minimal specification of the other equations in the system. In contrast, the full information methods treat the system as a whole and require a complete specification of all the equations.

The distinction between limited and full information methods is, in part, simply one of statistical efficiency. As is generally true in inference problems, the more that is known about the phenomena being studied, the more precisely the unknown parameters can be estimated with the available data. In an interdependent system of equations, information about the variables appearing in one equation can be used to get better estimates of the coefficients in other equations. Of course, there is a trade-off: full information methods are more efficient, but they are also more sensitive to specification error and more difficult to compute.

Statistical considerations are not, however, the only reason for distinguishing between limited and full information approaches. Models of the world do not come off the shelf. In any application, the choice of which variables to view as endogenous (i.e. explained by the model) and which to view as exogenous (explained outside the model) is up to the analyst. The interpretations given to the equations of the model and the specification of the functional forms are subject to considerable discretion. The limited information and full information distinction can be viewed not simply as one of statistical efficiency but one of modelling strategy.

The simultaneous equations model can be applied to a variety of economic situations. In each case, structural equations are interpreted in light of some hypothetical experiment that is postulated. In considering the logic of econometric model building and inference, it is useful to distinguish between two general classes of applications. On the one hand, there are applications where the basic economic question involves a single hypothetical experiment and the problem is to draw inferences about the parameters of a single autonomous structural equation. Other relationships are considered only as a means for learning about the given equation. On the other hand, there are applications where the basic economic question being asked involves in an essential way an interdependent system of experiments. The goal of the analysis is to understand the interaction of a set of autonomous equations.

An example may clarify the distinction. Consider the standard competitive supply demand model where price and quantity traded are determined by the interaction of consumer and producer behaviour. One can easily imagine situations where consumers are perfectly-competitive price takers and it would be useful to know the price elasticity of market demand. One might be tempted to use time-series data and regress quantity purchased on price (including perhaps other demand determinants like income and prices of substitutes as additional explanatory variables) and to interpret the estimated equation as a demand function. If it could plausibly be assumed that the omitted demand determinants constituting the error term were uncorrelated over the sample period with each of the included regressors, this interpretation might be justified. If, however, periods where the omitted factors lead to high demand are also the periods where price is high, then there will be simultaneous equations bias. In order to decide whether or not the regression of quantity on price will produce satisfactory estimates of the demand function, the mechanism determining movements in price must be examined. Even

though our interest is in the behaviour of consumers, we must consider other agents who influence price. In this case a model of producer behaviour is needed.

This example captures the essence of many econometric problems: we want to learn about a relationship defined in terms of a hypothetical isolated experiment but the data we have available were in fact generated from a more complex experiment. We are not particularly interested in studying the process that actually generated the data, except in so far it helps us to learn about the process we *wish* had generated the data. A simultaneous equations model is postulated simply to help us estimate a single equation of interest.

Some economic problems, however, are of a different sort. Again in the supply–demand set-up, suppose we are interested in learning how a sales tax will affect market price. If tax rates had varied over our sample period, a regression of market price on tax rate might be informative. If, however, there had been little or no tax rate variation, such a regression would be useless. But, in a correctly specified model, the effects of taxes can be deduced from knowledge of the structure of consumer and producer decision making in the absence of taxes. Under competition, for example, one needs only to know the slopes of the demand and supply curves. Thus, in order to predict the effect of a sales tax, one might wish to estimate the system of structural equations describing market equilibrium.

The distinction between these two situations can be summarized as follows: in the one case we are interested in a structural equation for its own sake; in the other case our interest is in the reduced-form of an interdependent system. If our concern is with a single equation, we might prefer to make few assumptions about the rest of the system and to estimate the needed parameters using limited information methods. If our concern is with improved reduced-form estimates, full-information approaches are natural since specification of the entire system is necessary in any case. A further discussion of these methodological issues can be found in Hood and Koopmans (1953, chs 1 and 6).

Limited Information Methods

Consider a single structural equation represented by

$$y = Z\alpha + u \tag{1}$$

where y is a T-dimensional (column) vector of observations on an endogenous variable, Z is α $T \times n$ matrix of observations on n explanatory variables, α is an n-dimensional parameter vector, and u is a T-dimensional vector of random errors. The components of α are given a causal interpretation in terms of some hypothetical experiment suggested by economic theory. For example, the first component might represent the effect on the outcome of the experiment of a unit change in one of the conditions, other things held constant. In our sample, however, other conditions varied across the T observation. The errors represent those conditions which are not accounted for by the explanatory variables and are assumed to have zero mean.

The key assumption underlying limited-information methods of inference is that we have data on K predetermined variables that are unrelated to the errors. That is, the error term for observation t is uncorrelated with each of the predetermined variables for that observation. The $T \times K$ matrix of observations on the predetermined variables is assumed to have rank K and is denoted by X. By assumption, then, $E(X'u)$ is the zero vector. Some of the explanatory variables may be predetermined and hence some columns of Z are also columns of X. The remaining explanatory variables are thought to be correlated with the error term and are considered as endogenous. Implicitly, Eq. (1) is viewed as part of a larger system explaining all the endogenous variables. The predetermined variables appearing in X but not in Z are assumed to be explanatory variables in some other structural equation. Exact specification of these other equations is not needed for limited information analysis.

In most approaches to estimating α it is assumed that nothing is known about the degree of correlation between u and the endogenous

F

components of Z. Instead, the analysis exploits the zero correlation between u and X. The simplest approach is the method of moments. Since $X'u$ has mean zero, a natural estimate of α is that vector a satisfying the vector equation $X'(y - Za) = 0$. This is a system of K linear equations in n unknowns. If K is less than n, the estimation method fails. If K equals n, the estimate is given by $(X'Z)^{-1}X'y$, as long as the inverse exists. The approach is often referred to as the method of instrumental variables and the columns of X are called instruments.

If K is greater than n, any n independent linear combinations of the columns of X can be used as instruments. For example, for any $n \times K$ matrix D, α can be estimated by

$$(D'X'Z')^{-1}D'X'y \tag{2}$$

as long as the inverse exists. *Often* D is chosen to be a selection matrix with each row containing zeros except for one unit element; that is, n out of the K predetermined variables are selected as instruments and the others are discarded. If Z contains no endogenous variables, it is a submatrix of X, least squares can then be interpreted as instrumental variables using the regressors as instruments.

The estimator (2) will have good sampling properties if the instruments are not only uncorrelated with the errors but also highly correlated with the explanatory variables. To maximize that correlation, a natural choice for D is the coefficient matrix from a linear regression of Z on X. The instruments are then the predicted values from that regression. These predicted values (or projections) can be written as NZ where N is the idempotent projection matrix $X(X'X)^{-1}X'$; the estimator becomes

$$(Z'NZ)^{-1}Z'Ny \tag{2'}$$

Because $N = NN$, the estimator (2) can be obtained by simply regressing y on the predicted values NZ. Hence, this particular instrumental variables estimator is commonly called *two-stage least squares.*

The two-stage least-squares estimator is readily seen to be the solution of the minimization problem

$$\min(y - Za)'N(y - Za). \tag{3}$$

As an alternative, it has been proposed to minimize the ratio

$$\frac{(y - Za)'N(y - Za)}{(y - Za)'M(y - Za)} \tag{4}$$

where $M = 1–N$ is also an idempotent projection matrix. This yields the *limited-information maximum-likelihood* estimator. That is, if the endogenous variables are assumed to be multivariate normal and independent from observation to observation, and if no variables are excluded a priori from the other equations in the system, maximization of the likelihood function is equivalent to minimizing the ratio (4). This maximum likelihood estimate is also an instrumental variable estimate of the form (2). Indeed, the matrix D turns out to be the maximum likelihood estimate of the population regression coefficients relating Z and X. Thus the solutions of (3) and (4) are both instrumental variable estimates. They differ only in how the reduced-form regression coefficients used for D are estimated.

The sampling distribution of the instrumental variable estimator depends, of course, on the choice of D. The endogenous variables in Z are necessarily random. Hence, the estimator behaves like the ratio of random variables; its moments and exact sampling distribution are difficult to derive even under the assumption of normality. However, large-sample approximations have been developed. The two-stage least-squares estimate and the limited information maximum-likelihood estimate have, to a first order of approximation, the same large-sample probability distribution. To that order of approximation, they are optimal in the sense that any other instrumental variable estimators based on X have asymptotic variances at least as large. The asymptotic approximations tend to be reasonably good when T is large compared with K. When $K - n$ is large, instrumental

variable estimates using a subset of the columns of X often outperform two-stage least squares. Further small-sample results are discussed by Fuller (1977).

Full Information Methods

Although limited-information methods like two-stage least squares can be applied to each equation of a simultaneous system, better results can usually be obtained by taking into account the other equations. Suppose the system consists of G linear structural equations in G endogenous variables. These equations contain K distinct predetermined variables which may be exogenous or values of endogenous variables at a previous time period. The crucial assumption is that each predetermined variable is uncorrelated with each structural error for the same observation.

Let $y_1,\ldots,y_G$ be T-dimensional column vectors of observations on the G endogenous variables. As before, the $T \times K$ matrix of observations on the predetermined variables is denoted by X and assumed to have rank K. The system is written as

$$y_i = Z_i\alpha_i + u_i (i = 1,\ldots,G) \tag{5}$$

where Z_i is the $T \times n_i$, matrix of observations on the explanatory variables, u_i is the error vector, and a_i is the parameter vector for equation i. Some of the columns of Z_i, are columns of X; the others are endogenous variables.

Again, estimates can be based on the method of moments. Consider the set of GK equations

$$X'(y_i - Z_i a_i) = 0 \qquad (i = 1,\ldots,G) \tag{6}$$

If, for any i, K is less than n_i the corresponding parameter α_i cannot be estimated; we shall suppose that any equation for which this is true has already been deleted from the system so that G is the number of equations whose parameters are estimable. If $n_i = K$ for all i, the solution to (6) is obtained by using limited information instrumental variables on each equation separately. If, for some i, $n_i < K$, the system (6) has more equations than unknowns. Again, linear combinations of the predetermined variables can be used as instruments. The optimal selection of weights, however, is more complicated than in the limited-information case and depends on the pattern of correlation among the structural errors.

If the structural errors are independent from observation to observation but are correlated across equations, we have the specification

$$E\left(u_i u_j'\right) = \sigma_{ij} I (i,j = 1,\ldots,G)$$

where the σ's are error covariances and I is a T-dimensional identity matrix. As a generalization of (3), consider the minimization problem

$$\min \sum_i \sum_j (y_i - Z_i a_i)' N \left(y_i - Z_j a_j\right) a^{ij} \tag{7}$$

where the σ^{ij} are elements of the inverse of the matrix $[\sigma_{ij}]$. For given σ's, the first-order conditions are

$$\sum_j Z_i' N \left(y_i - Z_j a_j\right) a^{ij} = 0 \,(i = 1,\ldots,G) \tag{8}$$

which are linear combinations of the equations in (6). It can be demonstrated that the solution to (8) is an instrumental variables estimator with asymptotically optimal weights. In practice, the σ's are unknown but can be estimated from the residuals of some preliminary fit. This approach to estimating the a's is called *three-stage least squares* since it involves least-squares calculations at three stages, first to obtain the projections NZ_j, again to obtain two-stage least-squares estimates of the σ's, and finally to solve the minimization problem (7). For details, see Zellner and Theil (1962).

If the structural errors are assumed to be normal, the likelihood function for the complete simultaneous equations system has a relatively simple expression in terms of the reduced-form parameters. However, since the reduced form is nonlinear in the structural parameters, analytic methods for maximizing the likelihood function are not available and iterative techniques are used instead. Just as in the limited-information case, the

maximum-likelihood estimator can be interpreted as an instrumental variables estimator. If in (8) the least-squares predicted values NZ_i are replaced by maximum-likelihood predictions and if the σ's are replaced by their maximum-likelihood estimates, the resulting solution is the (full-information) maximum-likelihood estimate of the α's. See Malinvaud (1970, ch. 19) for details.

At one time full-information methods (particularly those using maximum likelihood) were computationally very burdensome. Computer software was almost non-existent, rounding error was hard to control, and computer time was very expensive. Many econometric procedures became popular simply because they avoided these difficulties. Current computer technology is such that computational burden is no longer a practical constraint, at least for moderate-sized models. The more important constraints at the moment are the limited sample sizes compared with the number of parameters to be estimated and limited confidence we have in the orthogonality conditions that must be imposed to get any estimates at all.

See Also

- ▶ Econometrics
- ▶ Identification
- ▶ Instrumental Variables
- ▶ Simultaneous Equations Models
- ▶ Two-Stage Least Squares and the k-Class Estimator

Bibliography

Fuller, W. 1977. Some properties of a modification of the limited information estimator. *Econometrica* 45: 939–953.

Hood, W., and T. Koopmans, eds. 1953. *Studies in econometric method.* Cowles foundation monograph no. 14. New York: Wiley.

Malinvaud, E. 1970. *Statistical methods of econometrics*. 2nd ed. Amsterdam: North-Holland.

Zellner, A., and H. Theil. 1962. Three-stage least squares: Simultaneous estimation of simultaneous equations. *Econometrica* 30: 54–78.

Full Communism

P. J. D. Wiles

In Marx and Marxism, Full Communism is that final state of humanity in which productivity is higher than wants and everyone can help himself in the warehouses (not shops!). Since productivity cannot be unlimited, this entails that wants are limited: a direct contradiction to one of the basic propositions of Western economics. This is only possible because *wants have been reduced to needs*. Originally a governmental concept, needs are accepted as valid by each consumer, and internalized to become the new wants.

If wants are to fall below productivity, people must work seriously but voluntarily, that is *work too must become a need and so again a want*. The link between labour and reward is cut, so that everyone gets a 'dividend' and no one gets a wage, however much or little, well or ill, he or she works – and never mind at what job. Moreover that dividend must in total quantity correspond to the individual's consumption needs, so it is *nearly equal* for all people.

Since people would be 'well brought up', they would not help themselves to more than their 'need dividend' should they have the opportunity – for example, in the common mess hall or at the clothing warehouse. In more moderate versions large durables and housing are not offered in profusion without control, but *rationed*. However, the basic principle is not to ration, but to issue on demand, to a body of consumers too idealistic to 'break the bank'. Either way, *no money* is used inside the community. Moreover in the extreme version *nothing is scarce*. The lack of scarcity *removes the optimal allocation problem*, and causes the end of economics (if we accept that definition of it) as an intellectual subject.

Though allocations need no longer be optimal they must still be made, both of goods and of labour. The *state*, however, meaning the coercive organs of the governing class, in this case the proletariat, *has withered away*; so there is a big

question-mark over the nature of this allocating authority. At least, since economic scarcity has ceased, its yoke is light. On the other hand this authority must be conducting the propaganda that persuades everyone to internalize the new value system. Short on police power, the authority is long on spiritual power. It might, for instance, well be a Communist party without a security police.

In particular, however, unpopular labour, and labour threatening to convey political power to its performers (notably within the allocating authority), must both be *rotated*. Indeed, in extreme versions, all jobs are rotated, to relieve boredom and broaden human development. This is the (utterly impossible and now very embarrassing to Soviet scholars) abolition of the division of labour. This foolishness stems from Marx and Lenin's notion that advanced technology simplifies all labour.

We have only used the words 'utterly impossible' once, and we have presented the whole concept in ordinary Western language. This is partly because the kibbutz does embody Full Communism in practice, as indeed do most monasteries and nunneries. Elements of it are also included by other organizations such as cities under siege, countries immediately after Communist revolutions, and military forces. Perhaps above all the nuclear family, even the extended family, brings this utopia down to earth.

The kibbutz and the family, the former hardly Marxist, the latter originally scheduled to disappear under Full Communism, both illuminate the Marxist neglect of the *spiritual diseconomies of scale*. The altruism that we feel in not 'breaking the bank' with our consumption need not be very warm, but it must be there, if only as a sense of duty. The larger our community, the less warmth and eventually the less duty we feel. *Homo economicus* simply becomes an empirically more probable mode. But for Full Communism he must be altogether negated, at least on the consumption side. However generous a view we take of needs, only a very 'well-brought-up' population can reduce its wants to that, or indeed to any other than an infinitely high, level. In particular, while we can always want very little more than what we now have, it is almost impossible to want nothing more. So wants always grow, and are fed by *envy* and exceed needs by more and more.

It is a commonplace that the modern kibbutz cannot stop people consuming, but it can make people work. Work, after all, is in part natural. Up to a (very variable) point it is thought of as a duty and a pleasure. Deprivation of it is felt as painful, even when income is constant. *Homo economicus* explains work very badly, however large or small, rich or poor, capitalist or socialist, our community: he is already negated, in all systems.

Planning Under Full Communism

The kibbutz has a labour committee, which has the fairly simple task of drawing up a labour plan each week; and a consumption committee which, in the avowed presence of economic scarcity, adopts a mix of the following allocation instruments:

(i) Free supply; one just takes what one wants. This rule reigns, in respect of quantity but not quality, in the mess hall. Note that if there had been prices demand here would have been inelastic in respect to both price and income. Similarly when Russia went through its Full Communism post-revolutionary fit (June 1918–April 1921) local transport and postage were made uncompromisingly moneyless.
(ii) Rationed supply: housing and all durables, even clothing.
(iii) Pocket-money and actual prices: 'imported' luxuries such as cigarettes and sweets; coin-boxes such as telephones (also 'imported').

The pocket-money is of course divided equally, but the intrusion of money into utopia is viewed with grave misgiving. Not only is it bad in itself, but it leads to 'heterogeneous but equal' consumption. People receive unequal quantities of each thing, and this is supposed to give rise to envy, despite the overall equality of consumption volume. Another intrusion of 'money' is the use of shadow-prices by the labour committee. This is less bad in itself, but leads to narrow rationalistic

calculations, whereas Full Communism requires the broad sweep of 'policy' irrespective of mere economics.

Mutatis mutandis Communist governments take the same attitudes as kibbutzim. Of course, after their post-revolutionary fit they recognize that they are only in the 'socialist' transitional phase, in which only the enterprise and not the worker/consumer figures in the command plan; the latter is guided by prices and wage-rates. But they feel they should at least be tending the sprouts of the higher phase to come. To the shadow-price problem described above is added the fact that passive inter-enterprise wholesale prices exist in reality. These must, for accounting and bonus-formation purposes, be actually paid, but have no allocative function (the far smaller kibbutz needs no such thing). It would be convenient and rational to bring the passive prices into line with the shadow-price (which has an allocative function but is never paid). Perhaps such a society, in which there were at least no retail prices and instruments (i) and (ii) of consumption planning were used, could be called Full Communism.

The *official Marxist name* for Full Communism is 'Communism'; we have used the longer phrase for clarity. The first post-revolutionary phase is 'Socialism'. Marx describes this in his Critique of the Gotha Programme in very brief terms that correspond respectably to what the Soviet economy has become. Thus it is false that Marx left no post-revolutionary blueprint, but he certainly had a very foreshortened time path. He called the intermediate phase the 'Dictatorship of the Proletariat', and Full Communism, 'Socialism' or 'Communism' indifferently.

Full Communism and International Relations

A kibbutz is, in theoretical economics, a country. Hence our use above of the term 'imports'. People who leave it are 'emigrants', and so on. Like a communist country it uses 'foreign' money for its 'foreign' trade. But it is and is meant to be, even in high ideology, subject to the Israeli state, which is not about to wither away. However the Communist state is supposed to wither away, so who will guard its borders and administer migration and foreign trade? Some of these organs are by definition coercive. They can only wither away in a single world state – an irrefragable conclusion only lightly touched upon in Marxist writings.

See Also

- ▶ Anarchism
- ▶ Communism
- ▶ Socialism
- ▶ Utopias

Full Employment

G. D. N. Worswick

An expression which came into general use in economics after the Depression of the 1930s, full employment applies to industrially developed economies in which the majority of the economically active are the employees of firms or public authorities as wage and salary earners.

There has always been some unemployment in the course of development of capitalist economies and views have differed as to its causes and as to the extent to which it was a matter of public concern. In the first part of the twentieth century three principal strands of thought about unemployment can be distinguished. Firstly, the followers of Marx believed that cycles were an integral part of capitalist development and would lead to ever deepening crisis: the attempt to evade this by colonial expansion would only lead to conflict between imperialist powers. A second group of analysts paid particular attention to the measurement and dating of business cycles, distinguishing cycles of different periodicity, but they did not, as a rule, offer systematic theories. The third strand consisted of those economists

Full Employment, Table 1 Unemployed as a percentage of the total labour force

	France	Germany	Japan	Sweden	U.K.	U.S.A.
1900–1913	—	3	—	—	4.3	4.7
1920–1929	—	3.8	—	3.1	7.5	4.8
1930–1934	—	12.7	—	6.3	13.4	16.5
1935–1938	—	3.8	—	5.4	9.2	11.4
1950–1959	1.4	5.0	2.0	1.8	2.5	4.4
1960–1969	1.6	0.7	1.3	1.7	2.7	4.7
1970–1979	3.7	2.8	1.6	2.0	4.3	5.4
1980–1984	7.9	6.1	2.4	2.8	11.8	8.2

Sources: 1900–1979 A. Maddison, *Phases of Capitalist Development,* Oxford University Press, 1982. 1980–1984 OECD. *Main Economic Indicators,* Paris

F

who argued that in capitalist economies, if the forces of the market were left to work themselves out, there would always be a tendency towards an equilibrium, in modern parlance towards full employment.

Table 1 shows average rates of unemployment in six developed countries for various periods of the twentieth century. National estimates of unemployment are obtained either by sample survey or as the by-product of administration, such as a system of unemployment insurance. There are many problems in counting both the numbers unemployed and the labour force, whose ratio is to constitute the 'rate' of unemployment. There have been attempts to standardize rates obtained in different countries by different methods and over different periods. The figures in Table 1, taken from Maddison (1982) and OECD *Main Economic Indicators* are thought to be reasonably comparable. Only in two cases was it feasible to give estimates before World War I. We have four countries for the interwar years and all six after 1950. It will be seen that in the Depression years 1930–1934 the average rates of unemployment were far higher than in any earlier period in the twentieth century and that even in the later 1930s the rates remained abnormally high except in Germany.

The time was ripe for a theory which could account for the persistence of large-scale unemployment and it was provided by John Maynard Keynes in *The General Theory of Employment, Interest and Money* (1936), which the author himself said was all about 'my doctrine of full employment'. The self-equilibrating tendencies expounded by those whom (In the overlapping years 1975–1979 there are small discrepancies between Maddison and OECD for Germany and UK. The latest OECD figures were adjusted to be consistent with Maddison.) (In the overlapping years 1975–1979 there are small discrepancies between Maddison and OECD for Germany and UK. The latest OECD figures were adjusted to be consistent with Maddison.)

Keynes called 'classical' economists did not necessarily function in the manner prescribed for them and capitalist economies could get stuck with persistent unemployment. According to orthodox theory, unemployment should entail falling wages which would eliminate any 'involuntary' unemployment. Similarly, interest rates would fall, bringing about a recovery of investment. Keynes argued that money wages might be 'sticky', and even if they were not, falls in money wages would not entail corresponding falls in real wages, since prices would also fall. As to rates of interest, there was no guarantee that such falls as could occur would give a strong enough impetus to recovery. The analysis points clearly to the idea, which others developed more explicitly, that fiscal policy, that is, the adjustment of the budget balance between revenue and expenditure, could prove a more powerful lever to bring about full employment.

Within less than 10 years, the British wartime coalition government, in a famous White Paper, had accepted 'as one of their primary aims and responsibilities' the maintenance of 'a high and stable level of employment', and other governments, in Australia, Canada and Sweden, for instance, made similar affirmations. Article 55 of

the United Nations Charter called on members to promote 'higher standards of living, full employment, and conditions of economic and social progress and development'. This remarkable change in public policy cannot be attributed simply to the 'Keynesian Revolution' in economic thought. More powerful was the observation that twice in a generation full employment had only been realized in war. How far the new principles were responsible for the performance of economies in the postwar period is a disputed question. The facts are that for the 25 years after 1945 the growth rates of productivity in European countries were much higher, and the average levels of unemployment much lower than they had ever been. Fluctuations in output and employment were smaller than in the past. A group of OECD experts reporting in 1968 said that the results of using fiscal policy to maintain economic balance had been encouraging, though there was room for further improvement. In the United States, the government's attitude towards the new ideas was initially somewhat cooler. By its own past standards, productivity growth was not exceptional, and unemployment, though much lower than in the Depression, was much the same as in the 1920s and before 1914. The Keynesian battle was not truly joined in the USA until the 1960s. In the majority of countries, the era of exceptional growth and full employment came to an end in the early 1970s, since when longer spells of high unemployment have been experienced.

Full employment does not mean zero unemployment. There can be dislocations where large numbers of workers are displaced from their present employment, and time is needed before new workplaces can be created. This can happen at the end of a war, or following some major technological change. Apart from such special cases, regular allowance must be made for frictional and seasonal unemployment. Policy would not aim, therefore, at zero but at the elimination of unemployment attributable to demand deficiency. Governments targeting full employment would like to know the level of measured unemployment to which this corresponds. Three attempts to answer this question deserve mention. (1) The definition given by Beveridge (1944) was that the number of unemployed (U) should equal the number of unfilled vacancies (V). When U is very high, we would expect to find V low, and vice versa. If, over a number of fluctuations, U and V trace out a fairly stable downward sloping curve, we could pick the point on it where $U = V$ as indicating full employment. (2) Phillips (1958) claimed that for Britain there was a good statistical relationship between the level of unemployment and the rate of change of money wages. By choosing the level of unemployment delivering zero wage inflation, or when labour productivity was rising, the slightly higher level delivering zero price inflation, we could pinpoint full employment. (3) Friedman (1968) objected that in the long run there was no trade-off between unemployment and inflation: instead he argued that there was a 'natural' rate of unemployment, such that if the actual level was pushed below this, there would be not only inflation, but accelerating inflation. If this theory could be substantiated, one could choose the 'non-accelerating inflation rate of unemployment' (NAIRU) as the target. It is evident that the usefulness of each of the above approaches turns on the closeness and stability of the statistical relationship actually observed. Experience in different countries has varied, and the British evidence should be regarded as illustrative. For the period from the early 1950s to the later 1960s econometric analysis produced reasonably stable relationships for all three approaches, yielding estimates of the full employment level of unemployment of the order of 2–3%. But in the 1970s any stability of the Phillips curve crumbled, and estimates of NAIRU shot up from below two to over ten per cent, but without any clear indication of the institutional or structural changes which must have occurred to bring about so large a shift in so short a time. The UV relationship did not escape entirely unscathed either, but a plausible story can be told in terms of an outward shift of the UV curve. Brown (1985) reckoned that the United States, the United Kingdom and France suffered increases in the imperfections of the labour market in the period from the early 1960s to 1981 which might account in full employment ($U = V$) conditions for extra unemployment of 2% or less. It would seem that the substantial rises in unemployment, especially in

Europe, in the 1970s and 1980s can only be accounted in a smaller part by a rise in 'full employment' unemployment and that a greater part denotes a shortfall below it.

If the growth of output of developed economies after 1945 was exceptional, so also was the rate of price increase: in Britain, for example, such a sustained and substantial rise (3–4 % a year on average) had not been seen in peacetime for more than two centuries. Some countries had faster rises, but, in most cases, there was no clear sign of acceleration. A marked change of gear in price inflation occurred between the 1960s and the 1970s, precipitated by two large cost impulses. Around 1969 there was in many countries a distinct surge in wage increases which Phelps Brown (1983) has called 'the Hinge' and in 1973 there was the first of the great OPEC oil price rises. Confronted with these spontaneous boosts in costs, the authorities had to choose between allowing their consequences to be worked out within the bounds of the existing monetary and fiscal stance and adjusting that stance to accommodate them, which would mean that final prices would also jump. They began increasingly to opt for the former course. In doing so they received intellectual support from the first wave of the 'monetarist' counter-revolution against the now orthodox Keynesian demand management. Firstly, it was said that to push unemployment below the 'natural rate' would cause accelerating inflation. In any case, too little was known about the structure of the economy, in particular its time lags, for fine tuning to be a sensible policy. Better to adopt simple rules, such as fixed targets for the growth of the supply of money, which would keep inflation under control, and output and employment would adjust to the level indicated by the 'natural rate' of unemployment. Later developments in the new classical economics went further and denied altogether the possibility that governments, by loan financed expenditure, for instance, could effect lasting changes in employment. Instead, it was suggested, the only way to bring down unemployment was to reduce the monopoly power of trade unions, and to take other steps to free labour markets, such as abolishing minimum wage legislation and reducing unemployment benefit. Though not supported by any substantial body of evidence, these new ideas undoubtedly helped to persuade central banks to adopt fixed monetary targets, or rules, and after the second OPEC price rise in 1979, most governments followed restrictive monetary policies with more severe budgets. Calculations of 'constant employment' budget balances show a tightening equivalent to several percentage points of GNP in some cases, especially in Europe where unemployment rose considerably after 1980. On the other hand the United States broke ranks in 1983, allowing both actual and 'constant employment' deficits to rise, and it was the one major economy to experience falling unemployment.

If there is little evidence of a unique 'natural rate' of unemployment, it is nevertheless clear that to bring down a cost-induced inflation by demand restriction may involve high unemployment for a great many years. A wide range of 'income policies' has been attempted, and others canvassed, to secure that firms and workers would settle for lower prices and wages than they would seek if they were acting alone, provided others would do the same. It is unlikely that full employment of the kind experienced in Europe in the 1950s and 1960s could return without the aid of such policies. Throughout the great postwar expansion world trade grew at an unprecedented rate. Fixed exchange rates, with permission to change parities if needed, worked well enough for most countries to maintain their external balance. However, the Bretton Woods system crumbled and was succeeded by generally floating exchange rates, while at the same time controls over capital movements were being dismantled. Exchange rates came to be determined as much by capital movements as by trade, and they can diverge widely and for long periods from any level suggested by purchasing power parity. Thus full employment is also seen to depend increasingly on the joint action of all, or of a large number, of countries.

Employment policy has been linked with the welfare state in contradictory ways. On the one hand, higher unemployment is tolerated on the grounds that welfare provision mitigates the economic hardship involved: on the other hand, higher welfare costs are perceived as a growing burden on economies with high unemployment.

See Also

- ▶ Involuntary Unemployment
- ▶ Natural Rate of Unemployment
- ▶ Structural Unemployment
- ▶ Unemployment
- ▶ Wage Flexibility

Bibliography

Beveridge, W. 1944. *Full employment in a free society*. London: George Allen & Unwin.
Brown, A.J. 1985. *World inflation since 1950*. Cambridge: Cambridge University Press.
Friedman, M. 1968. The role of monetary policy. *American Economic Review* 58(1), March: 1–17.
Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.
Maddison, A. 1982. *Phases of capitalist development*. Oxford: Oxford University Press.
OECD. 1968. *Fiscal policy for a balanced economy*. Paris: Organization for Economic Cooperation and Development.
Phelps Brown, E.H. 1983. *The origins of trade union power*. Oxford: Clarendon Press.
Phillips, A.W. 1958. The relation between unemployment and the rate of change of money wage rates in the United Kingdom. *Economica* 25(November): 283–299.

Full Employment Budget Surplus

Terry Ward

The full or high employment budget surplus is a device for measuring fiscal stance and, specifically, a means of distinguishing the effects of discretionary budgetary policy on the economy from the autonomous effects on the budget of variations in economic activity. In other words, by estimating what public sector outlays, government revenue and, therefore, the budget balance would be, on the basis of current tax rates and expenditure programmes, the implications of policy action can potentially be isolated and the often misleading nature of changes in the actual budget balance kept in perspective.

Its origins lie in the recommendation made by the Committee for Economic Development in the United States that budgetary policy should be designed to 'yield a moderate surplus at high-employment national income' (Committee for Economic Development 1947, pp. 22–5). The purpose was essentially twofold: to try to make sure that automatic stabilizers – i.e. the tendency for the budget deficit to increase during a recession and to contract during a boom – were allowed to function without being nullified by policy action to bring the budget back to balance; and at the same time to limit the use of discretionary fiscal policy to stimulate economic activity and thereby to cause an unwanted and what was regarded as potentially damaging accumulation of public sector debt. It was a means therefore of keeping Keynesian demand management policies in bounds, which was important in a fiscally conservative country like the United States.

The concept was used most influentially by E. Cary Brown in 1956 in an analysis of the 1930s to demonstrate that federal deficits were caused predominantly by the depth of the recession rather than by lax fiscal policies. It was then taken up by a number of economists, Herbret Stein and Charles Schultze among others (Stein 1961 and Schultze 1961) to analyse policy in the economic downturn of 1960–61 and from then on has featured regularly in the US policy debate. Estimates have frequently been presented in the President's Budget documents, in annual reports of the Council of Economic Advisers, in Congressional Budget Office and in academic analyses of policy (such as Schultze et al. (1970–) and Pechman (1978–)).

In practice, the concept has been deployed both in periods of recession, in support of expansionary policies or as a warning against excessively deflationary ones, and in periods of economic upturn, to indicate the unsustainable nature of the budget deficits incurred as a means of shifting the economy out of recession. Given the process of fiscal policy-making in the United States, where any action taken is usually a compromise introduced only after a prolonged battle between the President and Congress, it is understandable that the reliance on fiscal stabilizers should be greater than in other countries and that the focus should be more on the longer term implications of present decisions. Though flawed, the full employment

budget surplus plays a useful role in this respect. It is relatively simple and straightforward to estimate – though there is often some disagreement over the rate of unemployment taken to represent full employment and the rate of growth required to maintain such a level – and therefore widely accepted as a meaningful if limited indicator of fiscal stance.

See Also

- ▶ Budgetary Policy
- ▶ Built-in Stabilizers
- ▶ Demand Management
- ▶ Fine Tuning
- ▶ Stabilization Policy

Bibliography

Brown, E.C. 1956. Fiscal policy in the thirties: A reappraisal. *American Economic Review* 46(5): 857–879.

Committee for Economic Development. 1947. *Taxes and the budget: A program for prosperity in a free economy*. Washington.

Pechman, J.A. 1978–. *Setting national priorities*. Washington: Brookings Institution.

Schultze, C.L. 1961. *Current economic situation and short-run outlook*, Hearings before the Joint Economic Committee, 86 Cong. 2 sess.. Washington, DC.

Schultze, C.L. et al. 1970–. *Setting national priorities*. Washington: Brookings Institution.

Stein, H. 1961. *January 1961 economic report of the President and the economic situation and outlook*, Hearings before the Joint Economic Committee, 87 Cong. 1 sess. Washington, DC.

Fullarton, John (1780–1849)

Roy Green

Keywords

Banking School; Convertibility; Currency School; Fullarton, J.; Law of reflux; Ricardo, D.; Tooke, T.

JEL Classifications

B31

John Fullarton shared at least one characteristic with his great predecessor, Ricardo: he also seemed, in the words of Lord Brougham, 'as if he had dropped from another planet'. Although Fullarton is described in the *Dictionary of National Biography* as a 'traveller and writer on the currency', travel occupied by far the greater proportion of his life, along with a keen interest in the world of art and literature. Yet the single published work on which his considerable reputation as an economist is based had an impact comparable with that of Ricardo's intervention in the Bullion Controversy at the turn of the century.

In his early twenties, Fullarton became a surgeon in India and found time to edit a Calcutta newspaper. There he subsequently made a fortune in banking and began the first of his extensive tours through 'our eastern possessions', as the *Dictionary of National Biography* endearingly calls them. On this tour, Fullarton collected vast amounts of information and made many notes of his observations, but these were never published. In 1823, having returned to England to live, he contributed articles to the *Quarterly Review* on the reform crisis; however, it was not long before he resumed his travels, this time around Britain and the continent in a coach specially fitted with a library. In 1833, as a Fellow of the Royal Asiatic Society, Fullarton went again to India, and, in the following year, to China; but his zeal evaporated along with his fortune as a result of the failure of his bankers, and he moved back to London permanently.

It was in 1844, during the passage of the Bank Charter Act through the House of Commons, that Fullarton published his major work, *On the Regulation of Currencies*, subtitled 'an examination of the principles on which it is proposed to restrict, within certain fixed limits, the future issues on credit of the Bank of England, and of the other banking establishments throughout the country'. It was immediately hailed as a formidable challenge to the Currency School orthodoxy, whose support for the Bank Charter Act had

overwhelmed Tooke's lonely opposition in the opening round of the 'currency-banking debate'. Indeed, according to Gregory, Fullarton's 'penetrating tract' was 'perhaps the most subtle and able production emanating from the Banking School' (introduction to Tooke, 1838/57, p. 81).

Fullarton's aim was a simple one: to bolster Tooke's case against what they both saw as ill-conceived banking legislation; in doing so, however, he not only improved its presentation, but also developed the theoretical basis of the argument in a number of important respects, taking the opportunity to lament the fact that 'Mr. Tooke himself has been exceedingly slow in following out his original conclusions on the subject of price to all their consequences' (1844, p. 18).

The Currency School had asserted that convertibility would not be a sufficient safeguard against the overissue of bank notes and their consequent depreciation; and that the quantity of notes in circulation would have to be regulated in accordance with the movement of bullion across the foreign exchanges. The response of Fullarton and the Banking School took three main lines. First, starting from the assumption that legal convertibility necessarily implied economic convertibility, they pointed out that any discrepancy between the note issue and a purely metallic system arose from the Currency School's erroneous theory of metallic circulation rather than from the supposed autonomy of the notes. Second, any effect on prices attributed to bank notes could not be denied to a range of financial assets excluded by the Currency School from their definition of money. Third, bank notes were in any case not money but credit, and therefore never could be overissued, though the credit structure as a whole might be extended beyond the limits of real accumulation by speculation. It was in this context that Fullarton developed the famous 'law of reflux', which he called 'the great regulating principle of the internal currency' (1844, p. 68).

Tooke, in turn, warmly welcomed Fullarton's analysis in the subsequent volume of his massive *History of Prices,* and gave some indication of the surprise he must have experienced upon its publication:

> [L]est his estimate of the value of my contributions to an extension of the knowledge of this subject, should be ascribed to the bias of friendship, I think it right to state that the distinguished author was unknown to me, except by name and reputation, till after the publication of his treatise, and that I had not the slightest knowledge of such a work being in preparation. (1838/57, vol. 4, pp. x–xi)

Tooke then paid Fullarton the compliment of quoting extensively from his work, repeatedly praising the 'wonderful clearness and vigour which distinguish his writings' (vol. 5, p. 537). Nor was Fullarton above self-promotion: it appears that he had a hand in a *Quarterly Review* article, 'The Financial Pressure', which saw the crisis of 1847 as confirming the warnings of 'Mr Fullarton's masterly treatise' (see Fetter 1965, p. 212).

It is certainly true that Fullarton's work 'enjoyed, in England and on the Continent, a persistent success such as few contributions to an ephemeral controversy have ever enjoyed' (Schumpeter 1954, p. 725). Marx, for example, included Fullarton among 'the best writers on money' (1867, p. 129); in his view, 'the economic literature worth mentioning since 1830 resolves itself mainly into a literature on currency, credit, and crises' (1894, pp. 492–3). Hilferding, too, drew heavily on Fullarton (Hilferding 1910); and even Keynes was impressed with his 'most interesting' contribution to monetary thought (Keynes 1936, p. 364 n.). Many of Fullarton's arguments later resurfaced in the Radcliffe Report of 1959, and are still today being 'rediscovered'. As Fullarton himself pointed out (1844, p. 5), 'this is a subject on which there never can be any efficient or immediate appeal to the public at large. It is a subject on which the progress of opinion always has been, and always must be, exceedingly slow.'

See Also

► Banking School, Currency School, Free Banking School

Selected Work

1844. *On the regulation of currencies*. London: John Murray.

Bibliography

Fetter, F.W. 1965. *Development of British monetary orthodoxy, 1797–1875*. Fairfield: Kelley. 1978.
Hilferding, R. 1910. *Finance capital: A study of the latest phase of capitalist development*. London: Routledge & Kegan Paul. 1981.
Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.
Marx, K. 1867. *Capital*, vol. 1. Moscow: Progress Publishers, 1959.
Marx, K. 1894. *Capital*, vol. 3. Moscow: Progress Publishers, 1971.
Radcliffe (Lord) (Chairman). 1959. *Committee on the working of the monetary system*. London: HMSO.
Schumpeter, J. 1954. *History of economic analysis*. London: George Allen & Unwin, 1955.
Tooke, T. 1838/57. *History of prices, and of the State of the Circulation from 1792 to 1856*, 6 vols. London: P.S. King & Son. 1928.

Functional Analysis

Leonid Kantorovich and Victor Polterovich

Abstract

A branch of mathematics mainly concerned with infinite-dimensional vector spaces and their maps, functional analysis is so called because elements (points) of certain important specific spaces are functions. The necessity of considering infinite-dimensional models arises in economics in many problems, including assessment of random effects in a situation with an infinite number of natural states; study of effects arising from a 'very large' number of participants; problems of spatial economics; study of economic development in continuous time, in particular, with due regard for lags; economic growth on an infinite time interval; and the influence of commodity differentiation on exchange processes.

Keywords

Competition models; Competitive equilibrium; Economic growth in the very long run; Extension principle; Fixed-point theorems; Functional analysis; Global analysis; Hyperplanes; Infinite-dimensional models; Kakutani theorem; Mathematical economics; Measure theory; Monopolistic competition; Openness principle; Product differentiation; Separation theorems; Spatial economics; Spectral analysis; Uniform boundedness principle

JEL Classifications

C6

Functional analysis is a branch of mathematics mainly concerned with infinite-dimensional vector spaces and their maps. Elements (points) of certain important specific spaces are functions, hence the term 'functional analysis'.

An important role in the development of functional analysis was played by set theory, abstract algebra and axiomatic geometry. General topology, measure theory, differential equations and some other branches of mathematics evolved in close contact with functional analysis, so that it is difficult to indicate where these disciplines end and functional analysis begins.

The fundamental ideas of functional analysis appeared at the turn of the 20th century; by the 1920s it had already evolved into an autonomous discipline. Among its founders were Banach, Fréchet, Hadamard, Hilbert, von Neumann, Riesz and Volterra.

The creation of functional analysis resulted in basic changes in the approach to many mathematical problems. The study of individual functions and equations was replaced by that of families of such objects. Abstract forms of investigation ensured a unified approach to questions which seemed distant at first glance; they were instrumental in finding more general, yet deeper and more concrete relationships.

From the outset, the development of functional analysis was stimulated by the intrinsic requirements of mathematics, as well as by applications, especially to quantum mechanics. Today the language of functional analysis is actually used in all of continuous mathematics. Its methods have become the foundation of a whole series of new

branches of research, both theoretical and applied, such as the theory of random processes, differential topology, dynamic systems, optimal control theory, mathematical programming, and so on. Functional methods penetrate deeper and deeper into theoretical physics and into different engineering disciplines. These methods find more and more widespread applications in mathematical economics.

Spaces studied in functional analysis usually belong to the class of linear (vector) topological spaces, that is, linear spaces supplied with a topology (a system of open sets and hence a notion of limit), for which the linear operations are continuous. A narrower class of spaces is metric vector spaces, for which distance between points is defined. The distance is given by a function (the metric, assigning a non-negative number to each pair of vectors) which possesses certain specific properties of ordinary distance. The topology in such spaces is naturally induced by the metric.

An important subclass of metric spaces is normed spaces, that is, linear spaces in which to each element x a non-negative number $\|x\|$, called the norm of x, is assigned, and the following conditions are satisfied:

(1) $\|x\| = 0$ if and only if $x = 0$;
(2) $\|\lambda x\| = |\lambda| \, \|x\|$ for any scalar λ (homogeneity);
(3) $\|x + y\| < \|x\| + \|y\|$ (triangle inequality).

The norm is an abstraction of the notion of 'vector length'. The function $d(x,y) = \|x - y\|$ is the metric in normed spaces. It is said that a sequence x_t of elements converges to the element x in the strong topology, if $\|x_t - x\| \to 0$ as $t \to \infty$. A normed space is said to be a Banach space if it is complete; this means that any of its fundamental sequences (that is, such that $\|x_t - x_s\| \to 0$ as $t, s \to \infty$) has a limit. Banach spaces often appear in applications.

A Banach space X is said to be a Hilbert space if it is supplied with a numerical function (x,y), called scalar product of vectors $x, y \in X$, related to the norm by the identity $\|x\|^2 = (x,x)$ and satisfying the conditions:

(1) (x, y) and (y, x) are complex conjugates (in particular, for real vector spaces, $(x, y) = (y,x)$);
(2) $(\lambda_1 x_1 + \lambda_2 x_2, y) = \lambda_1,(x_1 y_1) + \lambda_2(x_2, y)$;
(3) $(x, x) \geq 0$ and $(x, x) = 0$ only if $x = 0$.

The scalar product makes it possible to characterize the 'angle between vectors' and, in particular, to introduce the notion of orthogonal vector. As a result, the geometry of Hilbert spaces is close to Euclidean geometry.

Let us present some examples of specific spaces. The space $l_p (1 \leq p < \infty)$ of all numerical sequences $x = (\alpha_n)$ with the norm

$$\| x \| = \left(\sum_{n=1}^{\infty} | \alpha_n{}^p \right)^{1/p}$$

is a Banach space. For $p = 2$ it is a Hilbert space if the scalar product is defined by the formula

$$(x, y) = \sum_{n=1}^{\infty} \alpha_n \overline{\beta}_n, x = (\alpha_n), y = (\beta_n),$$

where $\overline{\beta}_n$ is the complex number conjugate to β_n. The space $L_2(a, b)$ of all real functions defined on the closed interval $[a, b]$, square integrable in the Lebesgue sense, is a Hilbert space (functions which differ on a set of zero measure are identified) if the scalar product is defined by the formula

$$(x, y) = \int_a^b x(t) y(t) \mathrm{d}t$$

$L_2(a, b)$ is a particular case of the Banach spaces $L_p (1 \leq p \leq \infty)$ of functions defined on so-called measure spaces. The theory of the spaces L_p is part of the foundations of probability theory, where the functions from L_p are interpreted as random variables. For $p \neq 2$ the spaces l_p and L_p are not Hilbert spaces.

Another important example is the Banach space $C(S)$ – the collection of all continuous scalar functions on the compact space S, with the norm

$$\| x \| = \max_{s \in S} \| x(s) \| .$$

All the spaces listed above are infinite dimensional, that is, contain an infinite subset of linearly independent vectors (the notion of linear independence here is the same as in linear algebra). A finite dimensional vector space may be transformed into a Banach space in many different ways by appropriate choices of norms, but the convergence in any norm will be equivalent to the coordinate one.

Although many facts of classical analysis can be generalized to Banach spaces, the infinite dimensional theory is essentially different from the finite dimensional one in many ways. One of the reasons is that a bounded sequence (with respect to norm) in a Banach space does not necessarily contain any fundamental subsequences and therefore may have no limiting points; such is the sequence l_n, $n = 1,2,...$ in l_2, whose nth element l_n is the vector all of whose coordinates are zero, except the nth, which equals 1.

A function from one space into another is often said to be an operator. Operators with scalar values are called functionals. The operators most thoroughly studied are the linear ones. An operator T from the vector space X to the vector space Y is called linear if

$$T(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1 T(x_1) + \lambda_2 T(x_2)$$

for all $x_1, x_2 \in X$ and arbitrary scalars λ_1, λ_2. In particular, the derivation and integration operations determine linear operators for appropriate choices of the spaces X, Y. If X and Y are finite dimensional, linear operators from X to Y are determined by matrices.

The theory of linear operators in Banach spaces is one of the most developed sections of functional analysis. It is a far-reaching generalization of linear algebra and, in particular, of matrix theory. However, the purely algebraic approach is insufficient in the infinite dimensional case. One of the reasons is the necessity of distinguishing continuous and discontinuous linear operators (continuity is not an algebraic notion), while for operators in finite dimensional space linearity implies continuity.

For a linear operator from one Banach space to another to be continuous, it is necessary and sufficient that it be bounded, that is, that it map bounded sets into bounded sets.

The set $B(x, y)$ of continuous linear operators from X to Y is a linear space with respect to the natural operations of addition and multiplication by scalars. This set becomes a Banach space if the norm $\| T \|$ of the operator T is defined by the formula

$$\| T \| = \sup_{\| x \| \leq 1} \| T(x) \| .$$

In the particular case when Y is the set of scalars, we get the Banach space X^* of all linear continuous functionals on X, which is called adjoint to X. The study of adjoint spaces is not only of intrinsic interest but is also needed to obtain deeper results about the initial space X.

The adjoint space of an n-dimensional space is also n-dimensional. The space adjoint to l_p coincides, in a certain sense, with the space l_q, where $1/q + 1/p = 1$(a similar statement holds for L_p). A complete description of linear continuous functionals has been obtained for many specific spaces. We only mention F. Riesz's famous theorem describing the general form of a linear continuous functional on the space $C(S)$ of continuous functions. In the particular case when S is the closed interval $[a, b]$ on the numerical line, any element $f \in C^*(a, b)$ can be represented in the form

$$f(x) = \int_a^b x(t) \mathrm{d}\varphi(t),$$

where φ is a function of bounded variation.

The operation of taking adjoint spaces can be iterated, yielding a sequence of Banach spaces X, X^*, X^{**}, ... each of which is adjoint to the previous one. Each vector $x \in X$ can be viewed as an element of the second adjoint space X by putting $x(f) = f(x)$ for any $f \in X^*$; the functional thus defined is linear, continuous and its norm coincides with $\|x\|$. If all the elements of X^{**} can be represented in this way, the initial Banach space X is called reflexive.

In certain aspects reflexive spaces have more resemblance to finite dimensional ones than do non-reflexive spaces.

F

A sequence x_n in a Banach space X is said to converge weakly to $x \in X$ if $f(x_n) \to f(x)$ as $n \to \infty$ for any functional $f \in X^*$. This definition implicitly supplies X with the weak topology which differs, as a rule, from the original one. The consideration of different versions of convergence on the same linear space and the study of their relationships is typical of functional analysis.

Among the numerous facts of Banach space it is customary to single out three theorems which, because of their importance and manifold applications, are known as the main principles of linear analysis.

The extension principle (Hahn–Banach Theorem) states that every continuous linear functional defined on a subspace of a normed space can be extended to the entire space, preserving norm. Using this principle it is possible to prove so-called separation theorems, which claim that under appropriate conditions two nonintersecting convex sets in a Banach space may be separated by a hyperplane, that is, a set of the form $\{x|f(x) = \alpha\}$, where f is a non-zero continuous linear functional and α is a scalar. Separation theorems make possible the wide use of geometric ideas in the study of Banach spaces.

The uniform boundedness principle (Banach–Steinhaus Theorem) states that a sequence of linear continuous operators $T_n \in B(X, Y)$ is pointwise convergent, that is, $T_n(x) \to T(x)$ as $n \to \infty$ for all $x \in X$ if and only if the two following conditions hold:

(1) Such a convergence takes place on a set of arguments whose linear envelope is dense in X;
(2) The norms of all the T_n are uniformly bounded with respect to n.

According to the openness principle (Banach Theorem), any continuous linear operator from one Banach space to another sends open sets into open sets.

The development of the theory of linear operators, especially at its initial stage, was stimulated by the problem of solving linear operator equations.

$$T(x) = y \tag{1}$$

where x, y are elements of infinite dimensional spaces.

The similarity between linear functionals and algebraic equations, previously noted for linear differential equations, turned out to be just as productive for integral equations, whose foundations were laid at the beginning of the century by Fredholm, Hilbert, Noether and Volterra.

An exhaustive theory has only been constructed for certain classes of equations (1). In particular, the case when $T = I + K$ where I is the identity operator and K is compact (that is, maps bounded sets into sets with compact closure) has been conclusively studied. Compact operators often appear in applications and are very similar to finite dimensional ones.

In the study of operator equations and in many applications of operator theory a leading role is played by the notion of spectrum. The spectrum of a continuous linear operator T defined in a complex Banach space is by definition the set of all scalars λ for which the operator $T - \lambda I$ has no inverse, that is, $T - \lambda I$ is either not injective (one-to-one) or not surjective (onto). Non-zero solutions of the equation $T(x) = \lambda x$ are called eigen-vectors of the operator T, while the values of λ for which such solutions exist are its eigen-values. All the eigenvalues are contained in the spectrum, but, unlike the finite dimensional case, the spectrum may also contain other values. A compact operator has a spectrum containing a finite or countable number of distinct numbers; in the latter case they converge to zero. Spectral analysis – the branch of functional analysis studying the properties of operator spectra – has achieved penetrating advances in the theory of Banach and operator algebras (Gelfand, von Neumann).

A linear operator T in Hilbert space is called self-adjoint if $(T(x), y) = (x, T(y))$ for all x, y. A compact self-adjoint operator has properties similar to that of a symmetric matrix; for example, there exists an orthonormal basis consisting of its eigen-vectors (Hilbert–Schmidt Theorem).

Among the branches of functional analysis beyond the framework of the theory of Banach spaces, the theory of distributions (or 'generalized functions'), initially developed (by Sobolev and Schwartz) as a rigorous foundation for formal

operations with δ-functions used in physics, should be mentioned.

In many theoretical and applied problems – in particular, in mathematical economics – it is necessary to consider semi-ordered vector spaces, characterized by the fact that some of their elements are involved in a comparison relation. The most important are those semi-ordered spaces for which every bounded (in the sense of the order relation) subset possesses a least upper bound. The foundations of the theory of such spaces were developed in the 1930s by Kantorovich and are called Kantorovich spaces (K-spaces). For example, the spaces l_p and L_p have a natural partial order relation: one sequence is greater than another, if all the coordinates of the first are greater than the corresponding coordinates of the second; the function x is greater than y if $x(t)$ is greater than $y(t)$ for almost all t. A somewhat wider class is constituted by vector lattices, in which the existence of l.u.b. is guaranteed only for finite sets. In semi-ordered spaces the notion of positive (not necessarily linear) operator can be introduced in a natural way; this notion has been used to generalize the theory of positive matrices.

Positive operators are an important class of maps studied in non-linear functional analysis. Another important class – the monotone operators – includes operators in Hilbert space satisfying the inequality

$$(T(x) - T(y), x - y \leq 0) \text{ for all } x, y.$$

A third example is that of contraction operators, i.e. operators such that

$$\| T(x) - T(y) \| < \alpha \| x - y \| \quad \text{for some } \alpha < 1.$$

For those (and some other) classes of nonlinear operators, conditions for the existence and uniqueness of operator equation solutions have been obtained in global terms. But, just as in classical analysis, the most universal means of studying nonlinear problems is the differential calculus. Many facts of classical differential calculus (in particular, Taylor expansions and the implicit function theorem) have been generalized to Banach spaces.

Among the main instruments of mathematical economics, convex analysis and fixed-point theorems should be noted. Both are in essence branches of functional analysis. The recent extremely rapid advances in convex analysis have been stimulated by the requirements of the theory of extremal problems in abstract spaces (mathematical programming and optimal control). A typical extremal problem is to find the maximum of the functional $f(x)$ defined on the subset G of the space X under the constraints $T(x) \geq 0, x \in$ G where T is an operator from X to a linear topological space Y supplied with the partial order $\geq$. As in the finite-dimensional situation, here the necessary and sufficient conditions for the existence of an extremum (under appropriate assumptions) may be stated in terms of saddle points of the Lagrange function

$$L(x, y^*) = f(x) + y^*(T(x)),$$

where the Lagrange multiplier y^* is an element of the space Y^* adjoint to Y In deducing this condition, separation theorems, the differential calculus and theorems on the representation of linear functionals play a fundamental role.

In order to solve functional equations and extremal problems in functional spaces, various computational procedures have been developed. In particular, generalizations of gradient methods and Newton's method have been obtained (the first results here are due to Kantorovich); the Newton–Kantorovich method also turned out to be a powerful means of proving existence and uniqueness of solutions. Another approach to computational problems is based on the approximation of the given functional equation by a simpler one. The application of functional analysis methods leads to a general theory of such approximation methods within whose framework the rate of convergence is studied and error estimates are given for a series of computational procedures.

In certain cases approximate solutions may be obtained by computer in analytic rather than numerical form ('deductive computations').

The necessity of considering infinite-dimensional models arises in economics in many problems, among which the following may be

distinguished: (1) assessment of random effects in a situation with an infinite number of natural states; (2) study of effects arising from a 'very large' number of participants (competition models); (3) problems of spatial economics; (4) study of economic development in continuous time, in particular, with due regard for lags; (5) economic growth on an infinite time interval; (6) influence of commodity differentiation on exchange processes. This list is not exhaustive.

As a rule, it is possible in principle to use a finite dimensional model and then pass to the limit if necessary. However, the 'infinite dimensional' statement of the problem is often easier to study because a more powerful analytic apparatus may be applied.

The concept of adjoint (dual) spaces mentioned above is of fundamental importance in economics. In a typical case the elements of the given space are interpreted as utilized and produced goods, while elements of the adjoint space (continuous linear functionals) are prices; the value of the functional on the given product vector determines its cost (expenditures, profits, and so on). Then semi-ordered vector spaces, expressing the 'greater than' relationship for certain pairs of expenditure and production vectors and taking into consideration the positivity of prices, turn out to be a natural instrument.

In the use of functional analysis methods, a very delicate question is that of choosing the functional space into which the model should be 'embedded'; it is closely related to the chosen estimate of economic and social values.

As an example let us consider a problem of type (5). In stating dynamical optimal planning problems considerable difficulties are involved in the choice of a plan horizon and objectives for the end of a planning period. However, in many cases the initial interval of the optimal trajectory depends very weakly on these parameters and is close to the corresponding interval of the optimal (in a certain sense) infinite trajectory. This is one of the reasons growth on an infinite time interval is worth studying.

For a wide class of models it is possible to show that any optimal trajectory is the result of maximizing integral profits calculated in appropriately chosen prices. An effective way of studying this question is the following. Let us embed the set of all admissible trajectories of economic growth (that is, trajectories satisfying technological and resource constraints) in an appropriate Banach space X so that the adjoint space X^* is interpreted as the space of prices; the value of a continuous linear functional on a vector $x \in X$ may be interpreted as the integral of the profits obtained in motion along the trajectory x. The set of trajectories which are better than the optimal one does not intersect the set of admissible trajectories. Under appropriate conditions these two sets may be separated by a hyperplane. The corresponding continuous linear functional will determine the required price trajectory. Using this approach, it is possible to investigate the relationship between competitive equilibrium and optimum for an infinite time interval.

Another example of productive application of functional analysis concerns the influence of commodity differentiation on market processes, a problem occupying an important place in the theory of monopolistic competition. In the simplest case, product differentiation is characterized by a scalar parameter assuming values in the closed interval $[a, b]$. Each consumer may choose any finite number of different goods (that is, a finite number of points t_i on the interval) and acquire them in arbitrary quantities x_i as long as he satisfies his budget restrictions for the given prices. It is natural to assume that the price $p(t)$ depends continuously on the characteristic of the product $t \in [a, b]$, i.e. $p(t) \in C(a, b)$. The result of a consumer's choice is a finite set of pairs x_i, t_i which determines a continuous linear functional in the price space $C(a, b)$ according to the rule

$$z(p) = \sum_i x_i p(t_i);$$

then $z \in C^*(a, b)$. But $C(a, b)$ can be identified with a subset of its second adjoint space (see above). Thus, as usual, price is a continuous linear functional of the space of collections of goods $C(a, b)$. The fact that this space is adjoint to a certain Banach space considerably facilitates its study, since adjoint spaces possess useful

topological properties. The analysis of models based on this construction yields conditions under which a market with differentiated commodities and 'small' participants, similar to contemporary competitive markets, ensures an optimal distribution of resources (Mas-Colell 1975).

The proof of the existence of competitive equilibrium in the finite dimensional case is based on fixed-point theorems. Several such theorems, including the Kakutani theorem, are also valid for Banach spaces; however, in this case their application becomes more difficult because of the essential trait of infinite dimensional spaces mentioned previously – the non-compactness of the unit sphere. Another trait of infinite dimensional spaces is that special conditions are required for the separability of non-intersecting convex sets. Both of these circumstances considerably complicate the study of economic models.

In discussing the economic applications of functional analysis, two other disciplines closely related to it – measure theory and global analysis – should be mentioned. The first is widely used in the study of probabilistic models, as well as in models with a continuum of participants or products (see Hildenbrand 1974; Mas-Colell 1975). Global analysis, introduced into mathematical economics by Debreu and Smale, allowed us to understand the deeper structures of the sets of equilibrium states and to advance to the solution of equilibrium stability problems (see Smale 1981).

Above we mentioned some applications of functional analysis to economics. In their turn, the problems of economics have influenced the development of mathematics. This is natural since economics is a vast field of research, differing in principle from those classical physical and mathematical disciplines on the basis of which functional analysis developed. The theory of systems of linear inequalities developed a hundred years later than the theory of linear equations, and precisely because of the needs of economics.

Another interesting and important example is the transportation problem, which was first studied under the name of mass shifting problem by Kantorovich in 1942. The metric introduced in its study (interpreted as the expenditures required to shift a unit mass) has found numerous applications in functional analysis and some other fields. Many mathematical problems from functional analysis originating in economics still await their solution. In particular, the functional equations describing macroeconomic dynamics taking into account the differentiation of funds according to their time of creation have not been exhaustively studied (for example, see Kantorovich et al. 1978). It can be expected that further advances in the mathematical analysis of economics will become an even more powerful source in the development of mathematical methods, including functional analysis.

See Also

- ▶ Calculus of Variations
- ▶ Non-Standard Analysis
- ▶ Pontryagin's Principle of Optimality
- ▶ Roos, Charles Frederick (1901–1958)

Bibliography

Dunford, N., and J.T. Schwartz. 1958. *Linear operators*. New York: Interscience Publishers.

Ekeland, I., and R. Temam. 1976. *Convex analysis and variational problems*, Studies in mathematics and its applications. Vol. 1. Amsterdam: North-Holland.

Hildenbrand, W. 1974. *Core and equilibria of a large economy*. Princeton: Princeton University Press.

Kantorovich, L.V., and G.P. Akilov. 1984. *Functional analysis*. London: Pergamon Press.

Kantorovich, L.V., V.I. Zhiyanov, and A.G. Khovansky. 1978. The principle of differential optimization as applied to a singleproduct dynamical economic model. *Sibirski matematicheskii zhurnal* 19: 1053–1064.

Kutateladze, S.S. 1983. *Foundations of functional analysis*. Novosibirsk: Nauka.

Mas-Colell, A. 1975. A model of equilibrium with differentiated commodities. *Journal of Mathematical Economics* 2: 263–295.

Schaefer, H.H. 1971. *Topological vector spaces*. New York: Springer.

Smale, S. 1981. Global analysis and economics. In *Handbook of mathematical economics*, ed. K.J. Arrow and M.D. Intriligator, vol. 1. Amsterdam: North-Holland.

Functional Central Limit Theorems

Werner Ploberger

Abstract

Functional limit theorems are generalizations of classical central limit theorems. They allow us not only to approximate the distributions of sums of random variables, but also describe their temporal evolution. The necessary mathematical concepts as well as some sufficient conditions for convergence to a random walk are discussed.

Keywords

Central limit theorems; Convergence; Functional limit theorems; General limit theorems; Gordin's th; Invariance principle; Likelihood; Lindeberg condition; Martingale differences; Random walk; Separability; Skorohod metric

JEL Classifications

C10

Central limit theorems guarantee that the distributions of properly normalized sums of certain random variables are approximately normal. In many cases, however, a more detailed analysis is necessary. When testing for structural constancy in models, we might be interested in the temporal evolution of our sums. So for random variables X_i we are interested in analysing the behaviour of

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{t} X_i \tag{1}$$

as a function of t for $t \leq N$. It is convenient to normalize the time, too, and consider for $0 \leq z \leq 1$

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{Tz} X_i. \tag{2}$$

Another popular application is the asymptotic behaviour of the empirical distribution function or its multivariate generalizations, though we will only briefly discuss it.

'Functional limit ths' are generalizations of the classical central limit theorem (CLT). Instead of analysing random variables with values in **R**, I deal with random variables in more general spaces. Here I discuss only one specific example, namely the analysis of the properly normalized partial sums of random variables. In order to do so, I will first sketch the necessary concepts concerning the topology of the spaces involved. In particular, I want to demonstrate the necessity of using spaces and metrics which, at the first glance, may not look that plausible. The results are well known and can be found in many textbooks. A classical reference is Billingsley (1999). Another introduction in this field, more geared towards econometricians, is Davidson (1994).

Foundations: Metric Spaces and Convergence in Distribution

A common framework, allowing us to formulate more general limit theorems, assumes that our 'random variables' take values in so-called 'Polish spaces', which are just metric spaces which are separable and complete. So let us assume that we have given such a space E, with a metric $d(.,.)$ on it, so that there exists a countable dense subset and that the space is complete (that is, every Cauchy sequence converges). Examples are the finite-dimensional spaces with the usual distance. The space $C[0,1]$ of all continuous functions from $[0,1]$ to **R** (the set of real numbers), endowed with the metric

$$d_M(x,y) = \max_{0 \leq t \leq 1} |x(t) - y(t)|. \tag{3}$$

Let us assume that we have random variables X_n, X with values in E: Then we define convergence in distribution of X_n to X

$$X_n \rightarrow^D X \tag{4}$$

if and only if for all bounded, continuous functions φ from E to **R**,

$$E\varphi(X_n) \to E\varphi(X). \tag{5}$$

We can easily see that, in the special case of the space E being the set $\mathbf{R}$, our definition here is a generalization of the familiar concept of convergence in distribution. An 'invariance principle' is simply a statement convergence in distribution of random variables in a complex space.

If we have given a statement like (4), then for continuous φ and large n we can approximate the distribution of $\varphi(X_n)$ by the distribution of $\varphi(X)$. As an example, assume our underlying space is $C[0,1]$ (defined above), and our distance is given by (3). Suppose we have $X_n \to^D X$. We can easily see that the functions attaching each $z \in C[0,1]$ $\max_{0 \le t \le 1} z(t)$ or $\int_0^1 z(t)^2 \mathrm{d}t$ are continuous with respect to our metric.

Hence we can immediately conclude that

$$\max_{0 \le t \le 1} X_n(t) \to \max_{0 \le t \le 1} X_n(t) \tag{6}$$

or

$$\int_0^1 X_n(t)^2 \mathrm{d}t \to \int_0^1 X_n(t)^2 \mathrm{d}t, \tag{7}$$

where '$\to$' stands for the usual convergence in distribution of real-valued random variables. Sometimes it is, however, burdensome to establish continuity for some functionals, or we might even be forced to consider discontinuous functionals. In this kind of situation the following theorem is helpful. Since we only work in separable, metric spaces a function ϕ defined on a general metric space E is continuous at a point $x \in E$ if for all $x_n \to x$ $\varphi(x_n) \to \varphi(x)$. Otherwise the φ is called discontinuous in x, and let $D\varphi$ be the set of all points where φ is discontinuous. Now assume we have some random elements X_n, X and

$$X_n \to^D X.$$

Then we have the following theorem.

Theorem 1 *Suppose that*

$$P([X \in D\varphi]) = 0.$$

Then

$$\varphi(X_n) \to^D \varphi(X).$$

If the discontinuities of φ are a null set with respect to the limiting distribution, the distributions of $\varphi(X_n)$ can be approximated better and better by the distribution of $\varphi(X)$.

In any case, the usefulness of functional limit theorems depends on the set of continuous functions associated with our space. On the one hand, a metric with 'many' continuous functions will allow us to establish many limiting relationships like (6) or (7). On the other hand, it will be harder to establish convergence, since we have to show the relation (5) for *more* functions ϕ. Hence we have to compromise.

The Space *D*[0,1]

The first and most important application of functional limit theorems is the analysis of partial sums. When dealing with normalized sums like (1), (2) we encounter the first problem: we can easily let the time t or z be a continuous variable. but then the sum (1),(2) is a discontinuous function. Hence we have to look at spaces more general than $C[0,1]$. One such space is the space $D[0,1]$, defined to be the space of all bounded functions f which have only 'jumps' as discontinuities: at every time z the limits to the right and left of f ($f(z+0)$ and $f(z-0)$ exist).

Next we have to define a distance between the functions f, g from $D[0,1]$. The first candidate, namely the supremum-norm (3), has the disadvantage that the corresponding space is not separable: consider for each $a \in (0,1)$ the functions f_a defined as

$$f_a(z) = \begin{cases} 0 & \text{if} \quad z < a \\ 1 & \text{if} \quad z \ge a. \end{cases} \tag{8}$$

Then we can easily see that in the supremum norm (3), the distance between f_a and f_b is equal to 1. Since we have more than countable real numbers in (0,1), we cannot have a countable dense subset.

A distance better suited to this space is the so-called Skorohod metric. Let us first define the set Λ to be the set of all functions from [0,1] to [0,1] which are monotonically increasing, continuous, and map 0 and 1 into 0 and 1, respectively. Then define

$$d_S(f,g) = \inf_{\lambda\in\Lambda}\sup_z\left(|f(z)-g(\lambda(z))|+|z-\lambda(z)|\right).$$

The Skorohod distance is related to the maximal distance. The main difference, however, is that we do not compare the functions f and g for the same values. The Skorohod metric allows us to 'bend' the argument a little. This rather small modification has enormous consequences. The corresponding space is separable: that is, there exists a countable dense subset. The metric itself is not complete (that is, there exist Cauchy sequences which do not converge). There exists, however, an equivalent metric (that is, a metric which determines the same open sets, neighbourhoods, convergent subsequences, continuous functions, ...) which is complete. Moreover, we can easily see that

$$d_S(f,g) \le d_M(x,y),$$

so convergence in the maximum distance implies convergence in the Skorohod metric.

The next question is the set of continuous functions. We can easily see that some of the usual candidates, like for example the functional mapping each f to $\sup_{0\le z\le 1} f(z)$, are continuous. The functional mapping f to $f(z)$, however, is for $0 < z < 1$ *not* continuous. Hence th 1 will come in handy.

The most important types of limiting processes will all have continuous trajectories. Hence, the class of functionals covered by th 1 contains all functionals which are continuous in $C[0,1]$. For establishing this continuity, we have an interesting criterion.

Theorem 2 *Suppose we have $f \in C[0,1]$, and $f_n \in D[0,1]$ so that $f_n \to f$ in the Skorohod metric. Then we have convergence in the supremum metric* (3)*, too.*

This result may explain the usefulness of $D[0,1]$. On the one hand, the metric on $D[0,1]$ is weak enough to allow for separability. This has, however, the drawback that it is hard to establish continuity of a function in the general case.

If the limiting random element lies with probability 1 in $C[0,1]$, however, it is easy to check the requirements of th 1 for a function $\varphi : D[0,1] \to \mathrm{R}$. One only has to show that $\varphi(f_n) \to \varphi(f)$ if $f_n \to f$ *uniformly*, which is much easier to handle.

Examples for Limit Theorems

In this section, I want to bring some examples of functional limit theorems. Together with the discussion above, they can be used as 'building blocks' for the derivation of general limit theorems.

The first functional limit theorem is one of the most important, namely, the functional limit theorem for martingale differences. This theorem is of utmost importance in many statistical applications: the scores of the conditional likelihood functions are martingale differences. Furthermore, the theorem is quite general. It only assumes a Lindeberg condition (which is quite similar to the case of the classical central limit th) and some kind of normalization condition. The role of the standard normal distribution is played by the 'standard random walk' W. W is a random element with values in $C[0,1]$ (that is, a random function) with the following properties:

- $W(0) = 0$.
- W is 'Gaussian'. All finite-dimensional marginal distributions $(W(z_0),\ldots,W(z_k))$ are Gaussian with expectation 0.
- The covariance of $W(z_1)$ and $W(z_2)$ is $\min(z_1, z_2)$.

A quite tedious but well known proof shows that there exists such a random element, and that its distribution (the induced probability measure on $C[0,1]$) is unique. Moreover, it is easy to show that W has all the properties associated with a random walk: its increments are independent from past values.

Theorem 3 (McLeish 1974): *Suppose we have given a triangular array of random variables* $X_{i,n}$. $1 \le i \le n$, *together with some adapted* σ*-algebras* $\boldsymbol{F}_{i,n}$ *so that*

$$E\left(X_{i,n}/\boldsymbol{F}_{i-1,n}\right) = 0.$$

Furthermore assume that the following two conditions are satisfied:

1. *The 'norming condition' is satisfied:*

$$\sum_{i \le nz} E\left(X_{i,n}^2/\boldsymbol{F}_{i-1,n}\right) \to z$$

uniformly in probability as $n \to \infty$.

2. *The* '*conditional Lindeberg condition*' *is fulfilled: for all* $\varepsilon > 0$

$$\sum_{i \le n} E\left(X_{i,n}^2 I_{\left[|X_{i,n}| > \varepsilon\right]}/\boldsymbol{F}_{i-1,n}\right) \to 0$$

in probability as $n \to \infty$.

Then let us introduce the random elements S_n, *defined by*

$$S_n(z) = \sum_{i \le nz} X_{i,n}$$

for $0 \le z \le 1$. *Then the* S_n *converge in distribution to a standard random walk W.*

Another important class of processes are stationary processes X_n, $n \in \boldsymbol{Z}$. In general, we will not even have a CLT. If, however, the conditional expectation of X_n given the $X_0, X_{-1}, \ldots$ decreases sufficiently fast, we will have a functional limit theorem, analogous to Gordin's theorem. Let us define the normalized, partial sums S_n by

$$S_n(z) = \frac{1}{\sqrt{n}} \sum_{i \le nz} X_i.$$

Furthermore we will use the L_2-norm of random variables: Define $\|X\| = \sqrt{EX^2}$.

Theorem 4 (Peligrad and Utev 2005): *Assume that we have given a stationary process* X_i *so that* $\sum \frac{1}{n} \|E(S_n(1)/X_0, X_{-1}, \ldots)\| < \infty$, *for all i*

$$\frac{1}{n} \sum X_t X_{t-1} \to E(X_t X_{t-i})$$

and

$$\sigma^2 = \sum_{i \in \boldsymbol{Z}} E(X_t X_{t-i}) < \infty.$$

Then

$$\frac{1}{\sigma} S_n$$

converges in distribution to a standard random walk W.

These two theorems should only act as illustrations for functional limit theorems. Especially for stationary processes, more general theorems are available. A good survey about recent results can be found in Merlevede et al. (2006).

Conclusion

This short introduction article should serve only as an introduction to functional limit theorems. Over the years, a rich theory has developed unifying many aspects of the limiting behaviour of functions of random variables. In particular, I would like to mention the limiting theorems for empirical distribution functions and their generalizations (see for example van der Vaart and Wellner 1996, for a survey, and Andrews and Pollard 1994, for dependent random variables). These results can be used to derive 'uniform' central limit theorems.

See Also

▶ Central Limit Theorems

Bibliography

Andrews, D.W.K., and D. Pollard. 1994. An introduction to functional central limit theorems for dependent stochastic processes. *Revue internationale de statistique* 62: 119–132.

Billingsley, P. 1999. *Convergence of probability measures*. 2nd ed. New York: Wiley-Interscience.

Davidson, J. 1994. *Stochastic limit theory: An introduction for econometricians*. Oxford: Oxford University Press.

McLeish, D.L. 1974. Dependent central limit theorems and invariance principles. *Annals of Probability* 2: 620–628.
Merlevede, F., M. Peligrad, and S. Utev. 2006. Recent advances in invariance principles for stationary sequences. *Probability Surveys* 3: 1–36.
Peligrad, M., and S. Utev. 2005. A new maximal inequality and invariance principle for stationary sequences. *Annals of Probability* 33: 789–815.
van der Vaart, A., and J.A. Wellner. 1996. *Weak convergence and empirical processes*. Berlin: Springer.

Functional Finance

David Colander

Abstract

The term 'functional finance' was created by Abba Lerner to contrast with sound finance. It involves making decisions about the deficit and the money supply with regard to their functionality, not some abstract moralistic premise. While it seems to play no role in the dynamic stochastic general equilibrium model prevalent in macroeconomics today, it does play a potential role in a more complex model where heterogeneous agents with limited information interact in a model with many different aggregate equilibria. Yet Lerner's functional finance theoretical model is far too simple to be acceptable, even as a rough guide for policy.

Keywords

Budget deficits; Coordination problems; Functional finance; General equilibrium; Great depression; Incomes policy; Inflation; Keynesianism; Lerner, A.; Macroeconomic externalities; Money supply; Multiple equilibria in macroeconomics; National debt; Optimal taxation; Sound finance; Stagflation

JEL Classifications

E0; B2

In the debate about how to pull economies out of the Great Depression, Abba Lerner created a steering wheel metaphor to contrast his 'economics of control' approach to policy with the then prevailing 'laissez-faire' policy. He argued that the laissez-faire approach was similar to driving a car without a steering wheel, the natural result of which was that the economy continually crashed, veering off the road first in one direction and then in another. It was time, he argued, for the government to adopt a Keynesian 'economics of control' approach in which the government used an explicit steering wheel – functional finance – to keep the economy running smoothly.

To complement that distinction between economics of control and laissez-faire, he contrasted the laissez-faire policy of sound finance with the economics-of-control policy of functional finance. Sound finance involved a set of rules – always balance the budget except in wartime, and do not increase the money supply at a rate greater than the growth rate of the economy. The problem, for Lerner (1944, 1951), was that these rules of sound finance were not analysed; they were simply accepted as being right. Lerner argued that, when governments understood how the macroeconomy actually operated, they would adopt an alternative 'functional finance' set of rules. Under the rules of functional finance, decisions about the deficit and the money supply would be made with regard to their functionality – their effect on the economy – and not with regard to some abstract moralistic premise that deficits, debt and expansionary monetary policy are inherently bad.

The Rules of Functional Finance

Functional finance consists of the following three rules (Lerner 1941).

1. The government shall maintain a reasonable level of demand at all times. If there is too little spending and, thus, excessive unemployment, the government shall reduce taxes or increase its own spending. If there is too much spending, the government shall prevent inflation by reducing its own expenditures or by increasing taxes.
2. By borrowing money when it wishes to raise the rate of interest, and by lending money or

repaying debt when it wishes to lower the rate of interest, the government shall maintain that rate of interest that induces the optimum amount of investment.

3. If either of the first two rules conflicts with the principles of 'sound finance', balancing the budget, or limiting the national debt, so much the worse for these principles. The government press shall print any money that may be needed to carry out rules 1 and 2.

In proposing these rules of functional finance, Lerner's purpose was to shift thinking about government finance from principles of sound finance that make sense for individuals – such as running a balanced budget – to functional finance principles that make sense for the aggregate economy. Functional finance principles used the budget balance as a steering wheel: deficits increased economic activity, surpluses decreased economic activity. The budget balance had these effects because, in the Keynesian model, government spending and taxing decisions directly affected levels of economic activity. These effects had to be considered because, in the aggregate, the secondary effects of spending decisions and savings decisions, which Lerner and I (Colander 1979) called macro externalities, had to be taken into account, whereas in individual decisions they did not.

Lerner's stark presentation of these rules of functional finance caused much stir in the 1940s and 1950s, when most Keynesians, including Keynes himself, were politically more circumspect about what came to be known as Keynesian ideas for government fiscal policy than they became in the 1960s (Colander and Landreth 1996). Lerner's rules specifically ruled out worrying about the size of a country's budget deficit or national debt.

In the 1950s and 1960s, Lerner's functional finance rules became both the basis of most textbook presentations of Keynesian economics and the basis of textbook macroeconomic policy discussions. It became what was generally considered Keynesian policy. This could occur because Keynes's *General Theory* contained almost no discussion of policy; it did not mention fiscal policy, and yet there were strong political forces pushing for its use. Thus, when 'Keynesian policy' was attacked in the late 1960s and early 1970s, it was primarily the idea of Lerner's policy of functional finance that most people were attacking (see Colander 1984, for a discussion).

That attack on 'Keynesian policy' intensified through the 1970s and 1980s, and by the 1990s textbook presentations of Keynesian policies had faded away. As they did so, so too did the concept of functional finance, and by the early 2000s few economists under the age of 50 had heard of it.

While the term 'functional finance' has disappeared from the macroeconomic textbooks, its influence continues among macro policy economists. The rhetoric of policy-oriented macro economists and their reaction to recessions is now quite different from what it was in pre-Keynesian times. When presenting fiscal policy to voters, governments are far less likely to talk about balanced budgets. Today, the potential benefits of government deficits in a recession are recognized. Similarly, policy-oriented macroeconomists discuss fiscal policy generally in terms of debtcarrying capacity such as represented by deficits as a percentage of GDP, not the need for a balanced budget, as was the case with sound finance. Even when a policy of functional finance is not used, the functional-finance role of fiscal policy is still seen as important since the expectation that government functional-finance policy will be adopted when crises occur can reassure agents and provide stability to the economy.

Why Functional Finance Lost Favour

Functional finance lost favour for a variety of reasons. First, Lerner's discussion of functional finance did not consider the politics of government finance; it assumed that the government could change taxes and spending according to the needs of the macroeconomy. In reality, both spending and taxing are difficult political issues, and the needs of politics generally trump the needs of stabilization. Second, the lags between recognition of a problem and implementation of a policy were significant, and the policy would often go

into effect long after the situation had changed. In Lerner's automobile metaphor, it was as if the steering wheel and the wheels were connected with a 30-second lag, and the windshield was opaque. Third, functional finance is built upon an assumption that the government knows what functional finance policy is best to follow –in inflationary times, increase taxes and decrease the money supply; in recessionary times, decrease taxes and increase the money supply. In the 1970s, when both inflation and recession occurred simultaneously, the functional finance rules seemed to give contradictory advice. These practical problems with implementing functional finance eliminated much, if not all, of the benefit of the steering wheel.

The reaction of Keynesian economists to the practical and informational problems was to limit the use of the deficit as a tool for fine-tuning the economy; the fiscal policy tool was a sledge, not a ball-peen hammer. The economics profession's reaction to stagflation was to accept a high rate of unemployment as the trigger for implementing an expansionary policy. Lerner did not follow the profession. His reaction to the stagflation problem was to argue that much inflation was not the result of excess demand but was instead what he called sellers' inflation. Sellers' inflation operated quite apart from demand pressures. Depending on how sellers' inflation was dealt with, there could be either high full employment or low full employment (Lerner 1972).

Lerner saw sellers' inflation as so important that, beginning in the 1960s, he changed his research programme to centre on finding cures for sellers' inflation. He developed a market-based incomes policy in which property rights in prices are established, and individuals have to buy the right to change prices from others who change their price in the opposite direction (Lerner and Colander 1980). Under a market-based incomes policy, rights in value-added prices would be tradable, so that any firm wanting to change its nominal price would have to make a trade with another firm that wanted to change its nominal price in the opposite direction. Thus, by law, the average price level would be constant, but relative prices would be free to change. With inflation controlled by such a plan, the rules of functional finance would once more become relevant (Colander 1979). Politically, in the early 2000s such policies had little chance of even being considered by governments and had faded from economists' radar screen.

Macro Theory and Functional Finance

It was not only the practical problems of functional finance that led to its demise. It was also that the profession essentially dropped the theoretical model upon which the concept was based. Functional finance was based on a coordination-failure model of macroeconomics –when individuals spent or saved, they did not take into account the effect of that decision on the aggregate level of spending; thus the economy needed some mechanism to internalize the spending complementarity and thereby determine the aggregate level of spending.

Today, among theoretical macroeconomists macro policy is thought of in a dynamic stochastic general equilibrium framework, and fiscal policy is discussed within an optimal taxation framework that assumes a representative agent is optimizing over a long-term horizon. The intuition behind such models is that the effect of any government deficit is mitigated by compensatory changes in the representative agent's spending decisions. This occurs because the agent will be responsible for paying off that deficit in the future. In the now prevalent modern macroeconomic theoretical approach, the possible existence of macro externalities is essentially ruled since the representative individual is assumed to take all the indirect effects of spending into account.

Assessment of Functional Finance

So what should one make of functional finance? My view is that, theoretically, it remains important. The fact that much modern macroeconomic theory does not allow for the possible existence of macro externalities is, in my view, a problem of modern macro theory, not a problem with functional finance. The probability that the unique

equilibrium, perfect rationality, perfect foresight, representative agent model underlying much of modern macroeconomics has much relevance to the real-world macro problems that we face is exceedingly small.

The macroeconomic theory problem seems more appropriately described as a coordination problem in which heterogeneous agents with limited information interact in a model in which many different aggregate equilibria are possible due to enormous strategic complementarities among agents. With multiple equilibria and coordination problems, there is no presumption of global optimality of the equilibrium chosen by the market. Everyone can know of the existence of a preferable equilibrium, but may not be able to achieve it by private actions. We can say something about that question only when we have a theory of equilibrium selection mechanisms. Currently we have none. Thus, in a multiple equilibrium economy with coordination failures, there should be no general presumption that the private economy, given its institutions, arrives at an equilibrium preferable to one achieved with government guidance.

That said, the functional finance theoretical model of Lerner is far too simple to be acceptable, even as a rough guide for policy. To say that individuals have limited information and do not fully take account of future effects of policy is not to say that they take no account of them. Private institutions develop which do precisely that, and any meaningful theoretical macro model must integrate such forward-looking private institutions into its structure. Doing so will involve highly complex models in which model selection by agents, agent interdependency, and social interaction by multiple agents are taken seriously. We are a long way from making such models tractable, so any formal macro model incorporating usable rules of functional finance is long in the future.

See Also

- ▶ Budget Deficits
- ▶ Cost-Push Inflation
- ▶ Keynesianism
- ▶ Multiple Equilibria in Macroeconomics

Bibliography

Colander, D. 1979. Rationality, expectations and functional finance. In *Essays in post Keynesian inflation*, ed. J. Gapinski. Cambridge: Ballinger.

Colander, D. 1984. Was Keynes a Lernerian? *Journal of Economic Literature* 22: 1572–1575.

Colander, D., and H. Landreth. 1996. *The coming of Keynes to America*. Cheltenham: Edward Elgar.

Lerner, A. 1941. The economic steering wheel. *University Review*, June 2–8.

Lerner, A. 1944. *The economics of control*. New York: Macmillan.

Lerner, A. 1951. *The economics of employment*. New York: McGraw Hill.

Lerner, A. 1972. *Flation*. New York: Penguin Books.

Lerner, A., and D. Colander. 1980. *MAP: A market anti-inflation plan*. New York: Harcourt Brace Jovanovich.

F

Fundamental Disequilibrium

D. E. Moggridge

The Articles of Agreement of the International Monetary Fund stipulate in Article IV(5)a that 'a member shall not propose a change in the par value of its currency except to correct a fundamental disequilibrium'.

The term itself was present from the earliest drafts of the American proposals for a postwar international monetary institution which noted that changes in exchange rates 'shall be made only when essential to correction of a fundamental disequilibrium' (Horsefield 1969, vol. III, p. 43). The term became part of an agreed Anglo-American text relating to exchange-rate changes on 15 September 1943, when the British suggested the form of words eventually embodied in the Articles of Agreement. At that time there was an attempt to define the considerations which the Fund should or should not take into account in determining whether such a disequilibrium existed. There were also some subsequent discussions of whether it would be possible to devise an 'objective test' by which the appropriateness of an exchange rate might be determined. These attempts to define fundamental disequilibrium were later dropped as impracticable.

As Harry White later remarked, 'It was felt . . . that the subject matter was so important, and the necessity for a crystallization of a harmonious view so essential, that it was best left for discussion and formulation by the Fund' (Dam 1982, p. 91).

Since its inauguration the Fund has never attempted to define the term. In 1946, when it was asked by the United Kingdom whether, as the government had committed itself to full employment, steps necessary to protect a member from unemployment of a chronic or persistent character would be considered measures to correct a fundamental disequilibrium, the Fund replied that, yes, such measures were among those necessary to correct a fundamental disequilibrium and that on each occasion when a member proposed a rate change to correct a fundamental disequilibrium the Fund was required to determine in the light of all relevant circumstances whether the change was necessary (Horsefield 1966, vol. III, p. 227). The matter came up again in 1948, when in connection with a French devaluation it was asked whether the Fund could object to a par-value change if in its opinion the change was insufficient to correct a fundamental disequilibrium. The Fund resolved the question by accepting that it could in principle object, but that in reaching a decision on any proposed exchange-rate change the member country 'should be given the benefit of any reasonable doubt' (ibid.). These matters rested until the redrafting of the Articles associated with the Jamaica Second Amendment of 1976. At that time, a par-value system like that of 1946–1973 was only a possible future system, but the notion of fundamental disequilibrium still remained – and remained undefined. Perhaps the last word should lie with the Bank for International Settlements, which noted in 1945 that the likely practical test of the notion would be that 'a disequilibrium which cannot be eliminated by any method other than an alteration of exchange rates must be regarded as fundamental' (1945, p. 109, n. 1).

See Also

- ► International Monetary Institutions
- ► International Monetary Policy

Bibliography

Bank for International Settlements. 1945. *Annual report*. Basle: Bank for International Settlements.

Dam, K.W. 1982. *The rules of the game: Reform and evolution in the international monetary system*. Chicago: University of Chicago Press.

de Vries, M.G. 1985. *The international monetary fund 1972–1978: Cooperation on trial*. 3 vols, Washington, DC: International Monetary Fund.

Horsefield, J.K. 1969. The international monetary fund 1945–1965: Twenty years of international cooperation. 3 vols, Washington, DC: International Monetary Fund.

Fungibility

Donald N. McCloskey

Fungibility is a central notion in economics, though often unnoticed and unnamed. It means merely 'substitutable', and is in origin a Latin legal term meaning 'such that any unit is substitutable for another' (from *fungor* meaning 'do, discharge'). A debt can be discharged with any money, not merely moneys from a particular account. The task of a low-level administrator is to make accounts fungible with each other, so that pencil money may be spent for office parties when required; the task of a high-level administrator is to prevent this. Mother cannot give money 'for' a new refrigerator: the gift merely raises the recipient's income. Likewise the World Bank rule that the items 'financed by' the Bank must attain a certain level of social return is pointless. The $100 million given to a government will be used anyway for the marginal project in the government's list; the project 'for which the money is given' can be claimed to be any intramarginal one.

Because demands for grain are fungible a cut in Soviet orders for American grain does not cause a one-for-one fall in demands on American suppliers. Because money is fungible the prospect of a government pension will reduce the incentive to save privately. The last, 'winning' points in a football game are in no coherent sense *the* winning points, since points are fungible. On the same grounds 'the reasons' for a decision are meaningless: criteria for the decision are fungible.

Fuoco, Francesco (1774–1841)

A. Quadrio-Curzio

Born in Migano (Naples) on 22 January 1774, Fuoco devoted almost all of his life to the study of political economy and was a member of the Scientific Academies of Naples, Turin and Palermo. He died in Naples on 2 April 1841.

His work can be set within the framework of the development of the contemporary Italian school of thought, and he reflects some of its typical subjectivistic features: the idea of necessity as the basis of the functioning of the economic system; the subjective evaluation of the value of goods; the idea of economic activity as the outcome of natural tendencies; and the idea of the 'public happiness' as a state of equilibrium. At the same time he can be considered atypical of his school in view of several theoretical and methodological contributions which place Fuoco among the followers of David Ricardo, both for his deductive reasoning and for the central role attributed to the theory of rent. The type of society from which he took his inspiration was, after all, that of industrial Lombardy and of its entrepreneurial middle class. Especially famous among his work was *La magia del credito svelata*, elaborated as a consequence of collaboration with the businessman Guiseppe De Welz.

Selected Works

1824. *La magia del credito svelata*, 2 vols. Naples.
1825–7. *Saggi economici. Pisa*. Anastatic reprinting, 2 vols, ed. Oscar Nuccio. Rome: Bizzarri. 1969.
1829a. *Introduzione allo studio della economia industriale. Principi di economia civile applicati all'uso della forze*. Naples: Tip. Trani. Reprinted in Rassegna monetaria. 1937.
1829b. *Le banche e l'industria*. Naples.

Bibliography

Anziani, V.M. 1978. La scuola classica in Italia: il caso di Francesco Fuoco. *Richerche Economiche* 32(1): 65–96.
Cossa, L. 1892. *Introduzione allo studio dell'economia politica*. Milan: Dizionario biografico universale. 1842. vol. II, Passigli.

Furtado, Celso (1920–2004)

Mauro Boianovsky

F

Abstract

Celso Furtado (1920–2004) was one of the most influential Latin American economists of the twentieth century. He was head of the development division of the United Nations Commission for Latin America in the 1950s, where he helped to formulate the structuralist approach to economics. His *Formação Economica do Brasil* (1959) is the classic interpretation of the economic history of Brazil. In 1961 he published a collection of essays about the notion of underdevelopment and development as interdependent phenomena. Furtado's last contribution was his careful discussion in the 1970s of the concept of cultural and economic dependence in underdeveloped countries.

Keywords

Balance of payments constraint; Balanced growth; Baran, P; Big push; Brazil; Centre–periphery system; Dependency theory; Furtado, C; Gerschenkron, A; Import substitution; Industrialization; Inflation; Kaldor, N; Latin American development; Lewis, W. A; Nurkse, R; Presbisch, R; Robinson, J; Rosenstein-Rodan, P; Rostow, W; Structuralism; Surplus; Underdevelopment

JEL Classification

B31

Celso Furtado was born on 26 July 1920 in Pombal (in the state of Paraiba, northeast of Brazil), and

died on 20 November 2004 in Rio de Janeiro. Together with the Argentinean Raúl Prebisch, Furtado was the most widely read and influential Latin American economist of the second half of the twentieth century. A prolific writer, he published more than 20 books on the economic history of Brazil and Latin America, and on the theory and policy of economic development, many of them translated into English, French and other languages.

He graduated at Universidade do Brasil (Rio) in 1944 and received his doctorate from the Sorbonne (Paris) in 1948; his thesis was about the Brazilian colonial economy. Maurice Byé was his supervisor, but it was François Perroux who impressed him most at the time. Upon his return to Brazil in that same year, Furtado was invited to join the staff of the new United Nations Economic Commission for Latin America (ECLA) in Santiago. From 1950 to 1957 he was head of the development division of ECLA. He then left Santiago to spend an academic year at Cambridge University working with Nicholas Kaldor and Joan Robinson with a Rockefeller Foundation scholarship. In 1958 he was appointed director of the Brazilian National Development Bank, where he conceived the project that led to the creation of SUDENE (Development Agency of the Northeast of Brazil) in 1959, of which Furtado was the first director (Hirschman 1963, chapter 1). In 1962 he also became Brazil's first minister of planning, charged with drafting a national economic plan, a position he held until 1963. Deprived of his political rights following the military coup in 1964, he left Brazil to take up appointments at American and European universities. Furtado went back to Paris and became the first foreign professor to be appointed at the Sorbonne, where he taught development economics from 1965 to 1985. After Brazil returned to democracy he was appointed Minister of Culture (1986–1988), and elected to the Brazilian Academy of Letters and to the Brazilian Academy of Sciences in 1997 and 2003 respectively. (Furtado's autobiography, originally published in three volumes between 1985 and 1991, was collected in 1997; the first volume, with recollections from the 1950s, his most productive period, was translated into French in 1987.)

Structuralism and Economic History

Together with Prebisch and other economists at ECLA in the 1950s, Furtado was one of the formulators of structuralism in Latin American economics. His main contributions can be found in two books, both available in English. In his 1961 volume on economic development, which collected essays written during the 1950s, Furtado provided the most elaborate exposition of the structuralist analysis in the literature at the time. In his 1959 classic *Formação Economica do Brasil,* written in Cambridge in 1957–1958 and based on Furtado (1950, 1952, 1954), the structuralist approach was applied for the first time to the interpretation of the economic history of a Latin American country, an exercise Furtado would expand to the whole region in his 1969 book. Furtado's methodological innovation was the use of historical investigation to identify factors that are specific to each structure through time: 'bring history near to economic analysis, get from the latter precise questions and find answers in history' (1997, vol. 1, p. 205). In *Formação* he pioneered the use of modern income analysis to deal with historical phenomena by introducing macroeconomic models into the analysis of each phase of Brazilian economic development from the sixteenth century to the 1950s (see also Furtado, 1963, for a brief account). Furtado's role in the historiography of the industrialization process of Brazil in particular and Latin American in general may be compared to Alexander Gerschenkron's well-known interpretation (1952) of the late industrialization of Russia and other continental countries. Like Gerschenkron, Furtado examined industrialization from the point of view of history. Both rejected Walt Rostow's (1960) view that the economic development of different countries goes through a succession of phases to which a single analytical framework can be applied.

The main feature of the 1959 book is the argument that the economic history of Brazil (and other Latin American countries as well) must be based on an open growth model with international trade treated as an endogenous variable, since these countries' economies evolved as suppliers of raw materials to the world market. Furthermore, the income-distribution profile is a main determinant

of the economic growth process through its effect on the level and structure of domestic demand in different historical phases. Furtado shows that throughout the four centuries from 1530 to 1930 the Brazilian economy depended on external demand to provide stimulus to higher productivity without previous capital accumulation, with three long-period cycles – sugar exports (1530–1650), gold mining (1700–80) and the expansion of the world market for coffee (1840–1930) – and intervening periods of relative stagnation. That phase came to an end in the economic crisis of 1929, when the collapse of export-commodity prices cut the country's import-purchasing power in half. According to Furtado, the policy adopted by the Brazilian government at the time to maintain the coffee price – that is, buying the unmarketable coffee and burning it – had the effect of an unwitting 'Keynesian' anti-cyclical deficit-financing policy. This contributed to maintaining domestic demand and, together with the diminished capacity to import, pushed up domestic prices of imported goods and stimulated investments in import-substituting industrial consumer goods. That process marked the beginning of a new phase in the development of Brazil, based on internal demand and import-substituting industrialization. Brazilian late industrialization – as compared with that of the United States – is explained in part by the differences between the productive structure of Brazil's export agriculture and the small agricultural properties in the English colonies of North America. The Brazilian internal market was much thinner due to the concentration of income and property, which served to maintain its stagnant colonial structure. Moreover, whereas the United States participated in the first wave of the Industrial Revolution as exporter of a key raw material (cotton), the main cause of the relative backwardness of the Brazilian economy in the first half of the nineteenth century, according to Furtado, was the damming up of its exports and the increase of the subsistence sector with lower productivity. Also in contrast with the late industrialization of continental European countries in the second half of the nineteenth century studied by Gerschenkron, the import-substitution process in Latin America did not lead to an intensive development of producer goods industries or changes in international trade (exports of manufactured goods and imports of raw materials). The evolution of trade patterns in Latin American countries during their industrialization after 1930 was quite the opposite: exports were still based on a few commodities and imports concentrated on goods whose production required huge investment and/or advanced technology.

The Concept of Economic Underdevelopment

It was in attempting to explain the backwardness of Brazil that Furtado hit upon the idea that underdevelopment and development are two interdependent phenomena which appear simultaneously as part of the evolution of industrial capitalism. The theme was elaborated in his 1961 book, where Furtado put forward concepts of economic underdevelopment and development that have been largely accepted in the literature. An underdeveloped structure is one in which 'full utilization of available capital is not a sufficient condition to complete absorption of the working force at a level of productivity corresponding to the technology prevailing in the dynamic sector of the economy' (1961b, p. 141). Underdeveloped economies (as distinct from simply backward ones) are hybrid structures characterized by technological heterogeneity of the various sectors. This comes from the historical fact that the import-substituting industrialization process in those economies led entrepreneurs to adopt a technology compatible with a cost and price structure similar to that prevailing abroad. Technology becomes, therefore, an independent variable in economies where industrialization is induced from outside. Whereas industrialization in underdeveloped economies was determined by demand, the formation process of capitalist European economies in the eighteenth and nineteenth centuries was dominated by supply factors, which led Furtado to define economic development as the introduction of new combinations of production factors which increase labour productivity. Underdevelopment is regarded as a permanent feature of the centre–periphery system, not as a stage on the road to development. Those ideas originally

appeared in an essay written as the first critical comment on Ragnar Nurkse's notion of 'balanced growth' (advanced in Nurkse's 1950 Rio lectures), where Furtado pointed out that the dynamics of demand (internal and external) in underdeveloped economies should be studied in tandem with the process of accumulation. According to Furtado, underdeveloped countries lack incentives to save (because of the consumer habits of higher-income groups), not to invest. The accumulation process should be examined from the point of view of changes in the process of generation, utilization and appropriation of the economic surplus, especially as affected by foreign trade. Furtado first developed these ideas in an essay originally written in Portuguese in 1955 (two years before Paul Baran made the concept of surplus a central notion of his own approach to development) and further elaborated it as part of a comment on Paul Rosenstein-Rodan's theory of 'big push', made at the International Economic Association conference on economic development held in Rio in 1957, and in his 1967 and 1980 textbooks.

Foreign Trade and Dependency

One of the main aspects of the industrialization process of Latin American countries, as discussed by Furtado in 1958 and 1960, is the persistent tendency towards balance of payment crises and inflationary pressures. Anticipating some elements of the two- gap model later developed by Chenery and Bruno (1962), Furtado showed in a two- sector model featuring a modern and a backward sector how balance of payment disequilibrium could constraint the economic growth process under the assumption that the coefficient of imports in the investment sector is larger than in the consumption sector, as is typically the case in underdeveloped countries. Such chronic disequilibrium has structural (not monetary) causes and may lead to the 'strangulation' of economic growth. Another obstacle to growth is that, after the end of the 'easy' phase of the substitution of imported consumer goods, as industrialization advances to the production of intermediate and capital goods the rate of profit falls because of the higher capital output ratio accompanied by increasing income concentration and lower aggregate demand. This was an essential element of Furtado's (1965; 1970) interpretation of the slowdown of economic growth in Latin America in the early 1960s, but, as the Brazilian economy recovered in the late 1960s and early 1970s, Furtado's stagnationist argument was criticized by economists in Brazil (see Tavares and Serra 1973). Furtado (1972, 1974, 1978) eventually concluded that, after the two earlier periods of economic growth – determined respectively by comparative advantages and import-substitution –- the Latin American economy had entered a new dynamic path in which consumption demand by high-income groups could under certain conditions become the leading factor of the system. This led him to explore in detail a theme that had often come up in his writings in the 1950s: dependency theory.

Furtado argued that underdeveloped economies feature cultural dependence, that is, consumption patterns are historically transplanted from developed countries by the upper strata of the underdeveloped areas as a result of their appropriation of the economic surplus generated through comparative advantages in foreign trade. Such modernized component of consumption brings dependence into the technological sphere by making it part of the production structure. Dependent structures are also dualistic systems with unlimited supply of labour at a subsistence wage, as first described by Furtado (1950) in his investigation of the dynamics of the labour market in Brazilian economic history. This is close to W. Arthur Lewis's (1954) classic model, but, in contrast with Lewis, Furtado's conclusion is that industrialization within a dualist dependent structure reproduces this dualism and does not bring about a homogeneous system with real wages increasing in tandem with the average productivity of the economy. The relationship between the centre and the periphery in the world economy is defined not just by the unequal sharing of the benefits of development and technical progress (as in Prebisch's terms-of-trade argument) but by dependence involving domination and control of access to modern technology by transnational corporations. In Furtado's view, economic growth does not entail economic development in dependent and reflex economies, since it implies an

aggravation of both external and internal exploitation, and thereby tends to make underdevelopment even more acute.

See Also

- ▶ Dependency
- ▶ Gerschenkron, Alexander (1904–1978)
- ▶ Prebisch, Raúl (1901–1986)
- ▶ Structural Change

Selected Works

1950. Características gerais da economia brasileira. *Revista Brasileira de Economia* 4: 7–37.

1952. Capital formation and economic development (trans: Cairncross, J.). In *The economics of underdevelopment,* ed. A. Agarwala and S. Singh. Oxford: Oxford University Press, 1958.

1954. *A Economia Brasileira.* Rio: A Noite.

1955. O desenvolvimento econômico – ensaio de interpretação histórico-analitica. *Econômica Brasileira* 1: 3–24.

1958. The external disequilibrium in the underdeveloped countries. *Indian Journal of Economics* 38: 403–410.

1959. *The economic growth of Brazil – A survey from colonial to modern times.* Trans. R. Aguiar and E. Drysdale. Berkeley: University of California Press, 1963.

1960. Industrialization and inflation (trans: Schwartz, P. and Henderson, E.). *International Economic Papers* 12(1967): 101–119.

1961a. Comments on Professor Rosenstein-Rodan's paper. In *Economic Development for Latin America,* ed. H. Ellis and H. Wallich. London: Macmillan.

1961b. *Development and underdevelopment – A structural view of the problems of developed and underdeveloped countries.* Trans. R. Aguiar and E. Drysdale. Berkeley: University of California Press, 1964.

1963. The development of Brazil. *Scientific American* 209: 208–220.

1965. Development and stagnation in Latin America: A structuralist approach. *Studies in Comparative International Development* 1: 159–175.

1967. *Théorie du développement économique.* Trans. A. Silva and J. Peffau. Paris: Presses Universitaires de France, 1970.

1969. *Economic development of Latin America – A survey from colonial times to the Cuban revolution.* Trans. S. Macedo. Cambridge: Cambridge University Press, 1970.

1970. *Obstacles to development in Latin America.* Trans. C. Ekker. New York: Anchor Books.

1971. Dependencia Externa y Teoria Economica. *El Trimestre Economico* 38: 335–349.

1972. *Analyse du 'modèle' brésilien.* Trans. E. Treves. Paris: Anthropos, 1974.

1974. *Le mythe du développement économique.* Trans. E. Treves. Paris: Anthropos, 1976.

1978. *Accumulation and development.* Trans. S. Macedo. Oxford: Martin Robertson, 1983.

1980. *Brève introduction au* développement. Trans. A. Ahmed. Paris: Publisud, 1989.

1985. *La fantaisie organisée.* Trans. E. Bailby. Paris: Publisud, 1987.

1987. Underdevelopment: To conform or reform. In *Pioneers in development – Second series*, ed. G. Meier. New York: Oxford University Press for the World Bank.

1997. *Obra Autobiografica de Celso Furtado,* 3 vols. São Paulo: Paz e Terra.

2000. Celso Furtado (born 1920). In *A biographical dictionary of dissenting economists,* 2nd edn, ed. P. Arestis and M. Sawyer. Cheltenham: Edward Elgar.

Bibliography

Arndt, H. 1987. *Economic development – The history of an idea*. Chicago: University of Chicago Press.

Baer, W. 1969. Furtado on development: A review essay. *Journal of Developing Areas* 3: 270–280.

Baran, P. 1957. *The political economy of growth*. New York: Monthly Review.

Bielschowsky, R. 1988. *Pensamento Economico Brasileiro – O ciclo ideologico do desenvolvimentismo*. Rio: IPEA.

Cardoso, E. 1981. The great depression and commodity-exporting LDCs: The case of Brazil. *Journal of Political Economy* 89: 1239–1250.

Chenery, H., and M. Bruno. 1962. Development alternatives in an open economy: The case of Israel. *Economic Journal* 72: 79–103.

F

Dobb, M. 1965. Review of C. Furtado, Development and underdevelopment – A structural view of the problems of developed and underdeveloped countries (1964). *Economica* 32: 460–461.
Furtado, C. 1952. Capital formation. *International Economic Papers* 1954(4): 124–144.
Gerschenkron, A. 1952. Economic backwardness in historical perspective. In *The progress of underdeveloped areas*, ed. B. Hoselitz. Chicago: University of Chicago Press.
Hirschman, A. 1963. *Journeys toward progress*. New York: The Twentieth Century Fund.
Hunt, D. 1989. *Economic theories of development: An analysis of competing paradigms*. New York: Harvester Wheatsheaf.
Kay, C. 1989. *Latin American theories of development and underdevelopment*. London: Routledge.
Lewis, W. 1954. Economic development with unlimited supplies of labour. *Manchester School* 22: 139–191.
Lowe, J. 1996. *Crafting the third world: Theorizing underdevelopment in Rumania and Brazil*. Stanford: Stanford University Press.
Mallorquin, C. 2005. *Celso Furtado – Um Retrato Intelectual*. Rio: Contraponto.
Mueller, H. 1963. Review of C. Furtado, The economic growth of Brazil – A survey from colonial to modern times (1963). *Journal of Economic History* 23: 359–360.
Nurkse, R. 1953. *Problems of capital formation in underdeveloped countries*. Oxford: Blackwell.
Prebisch, R. 1949. The economic development of Latin America and its principal problems. *Economic Bulletin for Latin America* 7(1962): 1–22.
Rostow, W. 1960. *The stages of economic growth: A - non-communist manifesto*. Cambridge: Cambridge University Press.
Szmrecsányi, T. 2005. The contributions of Celso Furtado (1920–2004) to development economics. *European Journal of the History of Economic Thought* 12: 691–702.
Tavares, M., and J. Serra. 1973. Beyond stagnation: A discussion on the nature of recent development in Brazil. In *Latin America: From dependence to revolution*, ed. J. Petras. New York: Wiley.

Futures Markets, Hedging and Speculation

David M. Newbery

Abstract

Futures markets provide partial income risk insurance to producers whose output is risky, but very effective insurance to commodity stockholders at remarkably low cost. Speculators absorb some of the risk but hedging appears to drive most commodity markets. The equilibrium futures price can be either below or above the (rationally) expected future price (backwardation or contango). The various effects futures markets can have on market and income stability are discussed. Rollover hedges can extend insurance from short-horizon contracts over longer periods.

Keywords

Arbitrage; Backwardation; Capital asset pricing model; Cobweb models; Commodity stabilization scheme; Contango; Electricity markets; Expectations; Forward contracts; Futures markets; Futures markets, hedging and speculation; Hedging; Income stability; Information aggregation; Information sharing; Liquidity; Portfolio choice; Price discovery; Price stability; Rational expectations; Risk aversion; Risk premium; Risk sharing; Rollover hedges; Speculation; Subjective probability; Vertical integration; Working, H.

JEL Classifications

G13

Futures markets for grain emerged in Chicago in the middle of the 19th century and spread rapidly to other commodities and centres. Forward contracts, in which two agents agree on the details of a transaction for delivery at a specified future date, must date back to the beginnings of commerce itself, but the distinctive feature of a futures market is that the contracts are standardized, transactions costs are minimized, and liquidity is high, so that contracts can be, and typically are, bought and sold many times during their lifetime, in contrast to most forward contracts. The standard explanation for the role of futures markets is that they help to spread and hence reduce risks, and to motivate the collection and dissemination of relevant information. Forward markets provide the same risk-sharing opportunities, but the greater transparency and liquidity of futures markets

makes the latter far more potent institutions for 'price discovery'.

The question of how well futures markets (and securities markets more generally) perform this role of collecting, aggregating and disseminating information is a large and important topic, best handled under the wider heading of 'information'. If we assume agents have rational expectations and share common information, then the price-discovery role of futures markets can be ignored and remaining issues of risk-sharing studied in isolation. In this case there is little conceptual difference between futures and forward markets, and we can concentrate attention on the two characteristic modes of behaviour exhibited by these markets – speculation and hedging.

Speculation and Hedging in Commodity Markets

Speculation is the purchase (or temporary sale) of goods for later resale (repurchase), rather than use, in the hope of profiting from the intervening price changes. In principle, any durable good could be the subject of speculative purchase, but, if carrying costs are high, or the good is illiquid, then the margin between the buying and selling price will be large, and speculation in that good will be normally be unattractive. Liquidity in this context means that there exists a perfect, or nearperfect, market in which the good can be sold immediately for a well-defined price, and this requirement severely limits the range of assets available for large-scale speculation. There are two types of assets –commodities traded on organized futures markets, and financial assets (bonds, shares) whose properties lend themselves particularly to speculation. Hedging, on the other hand, typically refers to a transaction on a futures markets undertaken to reduce the risks arising from some other risky activity, whether producing the commodity, storing it, or processing it for final sale.

Thus a risk-averse wheat farmer may hedge his future harvest by selling October wheat futures in January, in which case he is 'long' in actuals and 'short' in futures. A risk-averse miller who anticipates being short of wheat may hedge by buying futures now, in which case he will be a 'long' hedger. Speculators may be on the long or short end of any transaction, but in aggregate their position must offset any net imbalance in the long and short hedgers' positions.

It might appear from this that hedging consists in shifting the price risk onto the speculators in return for a risk premium. This view of speculation, advanced by Keynes (1923) and Hicks (1946), has been challenged by Working (1953, 1962), who denies any fundamental difference between the motivations of hedgers and those of speculators. One danger with looking exclusively at the price risk is that it ignores the more fundamental quantity risks that give rise to the price risks. Once this is appreciated, it is possible to formulate a simple theoretical model in which all agents are alike in attempting to maximize their expected utility but differ in the risks to which they are exposed, and these differences motivate trade on futures markets. While the activities of speculators are quite well defined, those of 'hedgers' are in general a mixture of insurance and speculation, as we shall see.

The simplest model of speculation and hedging has just two time periods. In the first period farmers plant their wheat, and the futures market opens. In the second period the wheat is harvested, sold, and the futures contracts expire. There are only three types of agents – farmers, who produce wheat but do not consume it; speculators, who neither produce nor consume wheat; and consumers, who neither produce wheat nor trade on futures markets. All agents are assumed to have beliefs about the relevant variables, which can be described by (subjective) probability distributions, and their behaviour is described by the theory of expected utility maximization. There are n farmers, and for the moment suppose that they have no choice over the amount of wheat to plant, but only over the size of their sales on the futures market. In the first period farmer i believes that his second period output will be $\tilde{q}_i$ (a random variable), and that the market clearing price will be $\tilde{p}^i$, also a random variable. In particular, he believes that $\tilde{q}_i$and $\tilde{p}^i$ are jointly normally distributed. The price of futures is f observable now, and he sells z_i futures, so that he believes his second period income will be

$$\tilde{y}_i = \tilde{p}^i \tilde{q}_i + z_i\left(f - \tilde{p}^i\right), \tag{1}$$

a random variable. The farmer's utility function exhibits constant absolute risk aversion, A_i, and takes the form $U^i(y) = -k_i \exp(-A_i \tilde{y})$, where $\tilde{y}$ is the random component of his income. (Any non-random components can be absorbed into the constant, k_i.) This particular form has the property that maximizing expected utility is equivalent to maximizing.

$$W = Ey - \frac{1}{2} A \operatorname{Var} y, \tag{2}$$

where Ey is the expected value of income, Var y is its variance, provided, as in the case here, that y is normally distributed. (These are the standard assumptions of the capital asset pricing model for portfolio choice, and can be viewed as second-order approximations to more general utility functions; see Newbery and Stiglitz 1981.) If Eq. 1 is substituted in (2), and if z_i can be positive (futures sales) or negative (purchases), then the value of z_i that maximizes W is

$$z_i = \frac{\operatorname{Cov}\left(\tilde{p}^i, \tilde{p}^i \tilde{q}_i\right)}{\operatorname{Var} \tilde{p}^i} - \frac{E\tilde{p}^i - f}{A_i \operatorname{Var} \tilde{p}^i}. \tag{3}$$

Speculator j has no risky production, so for him $\tilde{q}_j$ is zero, and the first terms in (1) and (3) vanish. Thus the second term in (3) can be identified as the speculative term, and is readily interpreted. The perceived riskiness of the futures contract is measured by $\operatorname{Var} \tilde{p}^i$ and the cost of this risk as $1/2 A_i \operatorname{Var} \tilde{p}^i$, The expected return to selling a futures contract is $f - Ep^i$. In order to persuade a risk-averse speculator to *buy* futures and accept the risk, the return to *selling* must be negative, hence f must be below the expected spot price, – a situation of *normal backwardation.* The first term in (3) is the pure hedging term, for if the futures market appears *unbiased* (that is, $f = E\tilde{p}^i$) then there is no expected speculative profit, and the only motive for trade is the income insurance offered by the price insurance. The quality of income insurance depends on how well income pq and price risks are correlated; that is, on the ratio of the covariance to the variance. If output is perfectly certain, then income and price are perfectly correlated, the first term will be equal to q_i, and the farmer would sell his entire crop on the futures market if he believed it to be unbiased. In general, though, he will not believe it to be unbiased, and he will wish to speculate in addition to hedging. His net futures trade will reflect the balance of the desire to insure and the returns to speculating.

The futures market clears, so that the sum of z_i across all participants must be zero, and this condition will yield a value for the futures price. What this implies for the value of f and its relation for the subsequent spot price, p, depends on beliefs, as well as preferences. If agents hold *rational expectations,* and have full information about the nature of all production and demand risks, then they will agree on the common values of the expected spot price, Ep, and its variance, Var p. In such a case the only motive for trading on the futures market is to share risk, and speculators will be willing to absorb some of the risk in return, on average, for some profit. If all farmers face perfectly correlated production risk, and if the coefficient of variation of output is σ_q, of price is σ_p, and the correlation coefficient between price and output is r, then market clearing on the futures market gives the bias as

$$\frac{Ep - f}{Ep} = \frac{\overline{Q}.Ep\sigma_p^2\left(1 + r\sigma_q/\sigma_p\right)}{\sum 1/A_i} \tag{4}$$

and a farmer's futures sales will be

$$\begin{aligned} \frac{z_i}{Eq_i} &= \beta_i\left(1 + r\sigma_q/\sigma_p\right), \\ \beta_i &\equiv 1 - \frac{\overline{Q}}{Eq_i A_i \sum_j 1/A_j}, \end{aligned} \tag{5}$$

where $\overline{Q} = \sum Eq_i$ is average total output (see Newbery and Stiglitz 1981, p. 186). Thus β_i is a measure of the extent to which the farmer is more risk-averse than the average (the term in A_i) and more exposed to risk $\left(\overline{q}_i/\overline{Q}\right)$. If there are n identical farmers and m identical speculators, all with the same coefficient of absolute risk aversion, A, then $\beta = m|(n + m)$. If there is no output

risk, so $\sigma_q = 0$, then, while a farmer would sell his entire crop forward on an unbiased futures market, here he would only sell a fraction β representing the fraction of the total risk which the speculators are willing to bear. If the only source of risk is supply variability, then $r = -1$, $\sigma_q/\sigma_p = \varepsilon$, the elasticity of demand, and the farmer will sell a fraction of his crop $\beta(1-\varepsilon)$ on the futures market, possibly negative.

What lesson can be drawn from this very simplified model? First, futures markets allow speculators to bear some of the farmer's risks. The more highly correlated income and price risks, the better the market is at insuring farmers, but in general it will provide only partial insurance. It is, however, much better suited to providing insurance to stockholders who store the commodity after the harvest until needed for consumption or processing, and it is not surprising that most hedging is done by stockholders rather than farmers. Second, the greater the agreement over the expected spot price, and the less risk-averse are the speculators, the smaller will be the average perceived bias, and the larger will be the fraction of hedging to speculative sales by producers (or stockholders). Third, the greater the degree of agreement on the expected spot price, the more will speculation be a response to the demand for hedging services. The greater the disagreement on the expected spot price, the more likely it is that speculation, in the form of gambling over the expected spot price, will dominate the market. In a masterly series of studies, Holbrook Working showed that most commodity futures markets depend primarily on hedging for their existence, that the size of the open interest follows closely the demand for hedging of seasonal storage, with speculators standing ready to assume the risks offered by the hedgers (Working 1962). The cost of these hedging services (that is, the return to the speculators) was quite remarkably small. Thus for cotton traders, the *gross* profit per dollar of sales over a sample of some 3,000 trades was 0.023 of one per cent with the traders making losses on 15 out of 43 trading days. (Net profits after paying commissions and expenses were substantially less; Working 1953).

The issue of bias turns out to be more complex than the simple Keynes–Hicks risk-premium view, for even in a bilateral market of farmers and speculators the bias can go either way. Once stockholders and processors are brought into the picture, the relative demands for long and short hedges will change yet again, and in turn influence the direction of speculation (long or short) and hence of the risk premium, or bias. Hirshleifer (1988) examines the determinants of bias in a market with primary producers subject to output risk (growers) and intermediate producers (processors). He finds that processors tend to hedge long, but, if transaction costs are low, there is a downward bias in futures prices (backwardation). If transaction costs are high, growers are differentially driven from the futures market, and could reverse the bias to contango.

Effect of Speculators on Stability

Several important questions can be asked about the role of speculators. Do they tend to destabilize the spot market and/or the futures market? Do they improve efficiency? Do they have adverse macroeconomic effects? To the layman the association of speculative activity with volatile markets is often taken as proof that speculators are the cause of the instability, though the body of informed opinion is that the volatility creates a demand for hedging or insurance, which is met by the willingness of speculators to bear the risk. It is hard to test the proposition that speculation is stabilizing, for speculative activity (notably, stockholding) can take place without futures markets. In practice, the usual question is: do futures markets, which, by lowering transaction costs, greatly facilitate speculative behaviour, improve the stability of the spot market? Even this question is not straightforward. Futures markets provide an incentive to collect information about the future market-clearing spot price, though, as often with information gathering, there are public-good problems associated with its use. Much theoretical effort has been devoted to the question of whether futures prices perfectly reveal the relevant information available to participants, and, if so, what incentives would remain for its collection. It now appears that, except in special cases, the

information is only partially revealed in the market, leaving incentives for its collection, but nevertheless improving the forecasts of otherwise uninformed traders. If so, and if the spot market is intrinsically volatile (because of variations in supply caused by weather, or demand caused by the trade cycle), then better forecasts of future spot prices will tend to elicit compensating supply responses – if prices are expected to be high tomorrow, then it will pay to produce more, and to carry more stocks forward, tending to reduce, or stabilize, price fluctuations. To the extent that futures markets reduce storage risks, storage becomes cheaper, and this will tend to stabilize supplies and prices directly. On the other hand, anticipated disturbances will have a more immediate effect on current prices, and will tend to make them more responsive to news. A frost in Brazil expected to affect next year's coffee production is likely to have a more rapid effect on current coffee prices in the presence of a futures market than in its absence. Nevertheless, it improves the efficiency of the current market if it does respond to this relevant information.

The clearest example of the stabilizing effect of futures market is provided by cobweb models, in which producers base current production decisions on last year's realized price, with consequent self-sustaining fluctuations in output without any exogenous shocks. If a futures market is set up, then producers initially planning to expand production in response to last year's high price, and selling futures, would cause the futures price to fall to the predicted spot price, and would lead them to revise their incorrect production plans, hence eliminating the cobweb and stabilizing the market.

Two other factors bear on the question of market stability. It is clear that much hinges on the nature of expectations. Speculation without hedging is a zero-sum game, and, if two speculators, each holding different views of the future price, $E\tilde{p}$ trade with each other, one will gain while the other will lose. If they are rational, and risk-averse, they should not be willing to engage in such swaps. On this view, speculators who are more successful at forecasting the future price will make money, and those who are less successful will lose, and be forced to leave the market, until only the good forecasters are left, and they make money only in the course of moving futures prices towards the forecast spot price. However, it is possible that a steady supply of less good speculators, who add noise to the system, lose money and exit, to be replaced by others. Their presence may worsen the predictive power of the futures price or, by increasing the returns to information gathering by the informed speculators, may actually improve the predictive power of the futures prices (Anderson 1984a; Kyle 1984). Depending on the direction of the net effect of uninformed speculators, the presence of a futures market (which provides them with the opportunity to gamble) may improve or worsen the efficiency of the spot market.

The other possibility is that futures markets will provide opportunities for market manipulation, by the better informed at the expense either of the less well informed (corners, squeezes) or of the larger at the expense of the smaller. It is easy to show that the futures price has an effect on production decisions by extending the model of Eq. 1 to allow producers to choose inputs. In the case of pure demand risk (no output uncertainty) it can be shown that the producer will base his production decisions solely on the future price. Large producers (Brazil for coffee, OPEC for oil, and so on) may then find it profitable to intervene in the futures market to influence the production decisions of their competitors in the spot market, and in extreme cases may find it profitable to increase price instability, though the extent to which this is feasible will be limited by the supply of and risk tolerance of other speculators in the futures market (Newbery 1984). This is true even if all agents hold rational expectations, and share full information (except about the actions of the large producers). If some agents use naive forecasting rules to guide their futures trading, and if these rules are known to other agents who possess market power, then it may pay the large rational agents to destabilize the price and exploit the irrationalities in the forecasting behaviour of the naive agents (Hart 1977).

Although speculation may stabilize prices, it is quite possible for it to make prices more unstable, even if all agents have equal information and hold

rational expectations. Compare two possible arrangements. In the first, futures markets are prohibited, the commodity is perishable, so there is no scope for speculative storage or speculation on the futures market. The commodity can be produced by two methods, one perfectly safe, the other risky, but on average more profitable (for example, two varieties of irrigated rice, one higher-yielding but susceptible to rust in certain weather conditions). Farmers allocate their land between the two production techniques but, in the absence of the futures market, find the risky technique relatively unattractive and so produce little. In the second arrangement, futures markets are permitted and speculators are willing to trade for a very low risk premium. Farmers are now able to sell the crop forward, and are therefore more willing to produce the risky crop, whose supply is very variable. Total supply variability increases, and hence the spot price becomes more variable.

It is quite possible that destabilizing speculation of this type yields higher potential social welfare, for yields are higher, if riskier, and the risks are borne at relatively low cost. It is also perfectly possible for speculation on a futures market to be stabilizing (by reducing the costs of storage and therefore improving arbitrage between crop years) and yet make everyone worse off (see, for example, Newbery and Stiglitz 1981). We now know that, if the market structure is incomplete, creating additional markets can make matters worse. Speculation, which creates a market in price risks, does not thereby complete the market structure because quantity risks may remain imperfectly insured. The reason is that the market in price risks causes changes in the market equilibrium which affects the degree to which the other risks (income and quantity risks) are effectively insured. In particular, if prices are stabilized but quantities remain unstable, incomes may be less stable than if prices were free to move in response to the quantity changes.

Finally, there remains the old Keynesian question of whether speculation which succeeds in stabilizing prices will exacerbate income fluctuations. The argument, due to Kaldor (1939), is straightforward. Speculators undertake or assume the risks for storage, which then responds to mismatches in supply and demand. These stocks, or inventories of goods, will fluctuate markedly and will have the same macroeconomic effect as fluctuations in investment, tending, through the multiplier, to have a magnified effect on national income. Whether these speculative stock movements are stabilizing or destabilizing then turns on whether they offset or amplify the fluctuations in income associated with the mismatch in demand and supply that caused the stock change. Kaldor's view was that stock changes caused by supply shocks would tend to stabilize total income, while those caused by demand stocks would be destabilizing, but much will depend on the commodity price elasticities of demand and the nature of the various transmission mechanisms, particularly the lag structure. Nevertheless, the OPEC oil shocks have demonstrated that commodity supply shocks can cause significant macroeconomic disturbances, while the increasing ease of currency speculation as restrictions are removed and transaction costs lowered has reawakened the fear that speculation may, in some cases, destabilize income and impose needless costs.

Commodity Stabilization Schemes and Longer-Term Insurance

At various times governments and international agencies have argued that primary commodity price variability is costly to vulnerable, often poor, primary exporters, and that therefore some form of commodity stabilization scheme should be implemented. Such schemes are often poorly designed to minimize the cost of reducing risk and have a doubtful record (Newbery and Stiglitz 1981). One might also expect that, in the presence of the kind of market failure suggested by this costly risk, alternative institutions might emerge to reduce risk, and that is indeed the case, with futures markets being the most obvious solution to commodity price risk. If primary exporting countries can hedge the export commodity price variability, then their risk will be reduced, and would seem to be eliminated if all the risk arose from price variability, with no variability in output. This would be true if there were no serial

correlation in prices from year to year, but, as Deaton and Laroque (1992) found, there is considerable serial correlation for the 24 commodities they studied over the period 1900-87. Their results suggest that about one-quarter of price shocks are permanent, that three-quarters or more of the price shock will persist for at least a year, and even after two years typically 60 per cent of the price shock will persist. If countries (or producers) hedge only for the coming year, their income will still vary considerably from year to year. If they could hedge for many years ahead this problem would be reduced.

Most futures markets extend only a relatively short period ahead and, even when they extend out several years, active trading and hence liquidity is mostly confined to the near-term future, measured in months rather than years. Apart from primary exporters having to deal with serial correlation (or persistence in price shocks), producers making large, irreversible sunk investment decisions (for example, in an oil refinery, offshore oil exploitation, LNG liquefaction and regasification facilities, aluminium smelters, nuclear power stations) would make better investment decisions knowing future prices (of inputs and outputs). They would be able to borrow more cheaply if risk were reduced by contracts or hedging, reducing the cost of capital-intensive products.

Liquid futures extending out ten years would clearly help, but are lacking. In their absence, companies may prefer to vertically integrate down the supply chain to provide an implicit (if partial) hedge. Electricity and gas liberalization has been premised on separating out natural monopoly pipes and wires from potentially competitive services supplied over the networks, regulating the former and creating wholesale and retail markets for the latter. Vertical unbundling (particularly of generation and transmission) appears critical to delivering the efficiency benefits of competition (Newbery 2005), but increases risk as wholesale electricity and fuel markets are so volatile. Forward and futures markets for electricity (and fuels such as gas) exist but *basis* risk (the difference between the price of the product traded and that of interest to the contractor) is high and markets are very illiquid. Vertical integration between generation and supply (or retailing) reduces spot price risk but makes the market less contestable.

Nevertheless, it is possible to use a sequence of short-term futures markets to hedge longer-term risks through a sequence of rollover hedges. Kletzer et al. (1992) show how to compute an *n*-year rollover hedge for a commodity with serially correlated price risk, no output risk but supply responsive to expected price. The way the rollover works is to sell more futures initially than needed for one-period hedging, and then use the surplus futures sales to finance the next year's futures transactions. This is not perfect, for the amount of hedging required next year will depend on production, and that will depend on the futures price prevailing next year, not as yet known. Consequently, despite the absence of production risk, future output cannot be perfectly hedged, and there remains some residual risk (as there would be if there were output risk). Nevertheless, because the costs of risk increase with the square of the deviation, reducing the risk by a given fraction reduces the cost of risk by more than that fraction and can be worthwhile. The further forward the hedge extends, the lower is the extra risk benefit provided, until the extra costs outweigh the benefit, so there is an optimal length of such a hedge.

The idea of using rollover hedging and portfolios of futures of different maturities to reduce risk has proved powerful both in theory and in hedging practice. Ross (1997) considers a world in which commodity prices are determined by many factors, and that, given enough different futures contracts and sufficiently precise knowledge of the underlying model determining prices, it would be possible to devise a perfect hedge, although in practice any such hedge would be imperfect. Neuberger (1999) develops this approach to identify an optimal hedging strategy using futures of different maturities and thus hedge long-term exposures with a combination of short-term futures. Neuberger tests his model on crude oil futures traded on NYMEX from 1986 to 1994. He asks how well one can hedge a forward commitment to deliver oil in five years' time using two futures contracts of not more than nine months to maturity. The annualized volatility of the five-year

contract is 26 per cent and that of the hedged portfolio is less than one per cent, with a hedge of short 2.89 seven-month contracts (of 1,000 barrels) and long 3.93 nine-month contracts, for each contract to deliver in five years' time. In a model in which a trader wishes to hedge for delivery in 36 months' time, if the portfolio is balanced monthly, 488 contracts are traded per contract delivered, although this can be cut to fewer than 60 with bimonthly rebalancing (and at lower risk).

The fact that rollover hedges allow one to reduce risk over a longer time horizon than the duration of current futures offered in the market has a number of interesting implications. It can explain why near-term futures are more popular and liquid than longer-term contracts, for they may provide a substitute for the latter at lower cost. It also explains why the volume of liquid futures can so greatly exceed the underlying physical trade, often by factors of 10–20. Rollovers require both a greater ratio of futures to physicals and a higher rate of trading to rebalance the portfolio over time, contributing to volume, liquidity and hence cost reduction.

Rollovers are, however, not perfect, and they may tempt traders to take imprudently large risks. One such famous case was the near-bankruptcy of Metallgesellschaft (MG), whose losses were estimated at DM 4 billion and whose survival was ensured only by a major rescue operation (Wahrenburg 1996). At one time MG was reportedly holding short-term positions equivalent to 160 million barrels of oil or 80 times the daily output of Kuwait (Hilliard 1999).

The case became celebrated as a test of whether MG had adopted a sound or imperfect hedging strategy. Some writers such as Culp and Miller (1995) argued that MG was following 'a textbook hedging strategy which was not properly understood by MG's supervisory board and house banks' (Wahrenburg 1996, S29). Others, such as Edwards and Canter (1995), Mello and Parsons (1995), and Verleger (1999) argue that MG was excessively exposed in the wrong products. Wahrenburg argues that the MG's hedging strategy could indeed significantly reduce risk, but not completely, and that MG's equity capital was insufficient to cover the remaining risk.

See Also

- ▶ Arbitrage
- ▶ Hedging
- ▶ Information Aggregation and Prices
- ▶ Options
- ▶ Options (New Perspectives)
- ▶ Present Value

Bibliography

Anderson, R. 1984a. The industrial organization of futures markets: A survey. Ch. 1 of Anderson (1984b).

Anderson, R., ed. 1984b. *The industrial organization of futures markets*. Lexington: Lexington Books.

Culp, C., and M. Miller. 1995. Metallgesellschaft and the economics of synthetic storage. *Journal of Applied Corporate Finance* 7: 6–21.

Deaton, A., and G. Laroque. 1992. On the behaviour of commodity prices. *Review of Economic Studies* 59: 1–24.

Edwards, F., and M. Canter. 1995. The collapse of Metallgesellschaft: Unhedgeable risks, poor hedging strategy or just bad luck? *Journal of Futures Markets* 15: 211–264.

Hart, O. 1977. On the profitability of speculation. *Quarterly Journal of Economics* 91: 57–97.

Hicks, J. 1946. *Value and capital*. 2nd ed. Oxford: Oxford University Press.

Hilliard, J. 1999. Analytics underlying the Metallgesellschaft hedge: Short term futures in a multiperiod environment. *Review of Quantitative Finance and Accounting* 12: 195–219.

Hirshleifer, D. 1988. Risk, futures pricing, and the organization of production in commodity markets. *Journal of Political Economy* 96: 1206–1220.

Kaldor, N. 1939. Speculation and economic stability. *Review of Economic Studies* 7 (1): 1–27. Reprinted in N. Kaldor, Essays on Economic Stability and Growth. London: Duckworth, 1960.

Keynes, J.M. 1923. Some aspects of commodity markets. Manchester Guardian Commercial, Reconstruction Supplement 29, March.

Kletzer, K., D. Newbery, and B. Wright. 1992. Smoothing primary exporters' price risks: Bonds, futures, options and insurance. *Oxford Economic Papers* 44: 641–671.

Kyle, A., 1984. A theory of futures market manipulation. Ch. 5 of Anderson (1984b).

Mello, A., and J. Parsons. 1995. Maturity structure of a hedge matters: Lessons from the Metallgesellschaft debacle. *Journal of Applied Corporate Finance* 8: 86–105.

Neuberger, A. 1999. Hedging long-term exposures with multiple short-term futures contracts. *Review of Financial Studies* 12: 429–459.

Newbery, D. 1984. The manipulation of futures markets by a dominant producer. Ch. 2 of Anderson (1984b).

F

Newbery, D. 2005. Electricity liberalization in Britain: The quest for a satisfactory wholesale market design. *Energy Journal* 26: 43–70. Special Issue on European Electricity Liberalisation, ed. D. Newbery.

Newbery, D., and J. Stiglitz. 1981. *The theory of commodity price stabilization*. Oxford: Clarendon Press.

Ross, S. 1997. Hedging long-run commitments: Exercises in incomplete market pricing. *Economic Notes* 26: 385–419.

Verleger, P. 1999. Was Metallgesellschaft's use of petroleum futures part of a rational corporate strategy? *Journal of Energy Finance and Development* 4: 89–115.

Wahrenburg, M. 1996. Hedging oil price risk: Lessons from Metallgesellschaft. *Chicago Board of Trade Research Symposium Proceedings* 2: S29–S47.

Working, H. 1953. Futures trading and hedging. *American Economic Review* 43: 314–343.

Working, H. 1962. New concepts concerning futures markets and prices. *American Economic Review* 52: 432–459.

Futures Trading

H. S. Houthakker

The object of futures trading is the *futures contract*, which may be defined as a highly standardized forward contract. Although the terms 'forward' and 'futures' are often used interchangeably in the older literature, the distinction is essential to the understanding of futures trading. Forward contracts are widely used; thus an agreement in which an automobile dealer undertakes to deliver a car of a specified make, type and colour to a customer at some later date is a forward contract; so is an employment contract, in which the employee promises to perform specified services during a certain period of time. Because forward contracts are typically quite specific, the employee in the last example cannot substitute another worker for himself without the employer's consent. Futures contracts, by contrast, exist only for a limited number of commodities and financial instruments, and are used only by a relatively small number of firms and individuals.

Futures contracts are of two types. The traditional contract provides for actual delivery of the underlying merchandise or financial instruments. In the early 1980s contracts with 'cash settlement' were introduced; they are settled not by delivery but by calculating traders' gains and losses from a known price, for instance an index of equity prices. Cash settlement is inherently simpler than delivery, but it is of limited application because in most markets there is no single price that could be used for this calculation. The following discussion focuses on futures contracts with delivery, though most of it also applies to cash-settlement contracts.

The standardization characteristic of futures contracts generally involves five elements: (1) *Quantity*: buyers and sellers can deal only in lots of fixed size, for instance 5000 bushels of wheat or bonds with a face value of $100,000; of course they can buy or sell any number of such lots. (2) *Quality*: the commodity or instrument is usually not completely specified but can be anywhere in a range (e.g. all wheat of certain grades, or all government bonds maturing within a certain interval). (3) *Delivery time*: the lot can be delivered at any time within a specified period, say a month. In most markets only contracts for selected delivery months are traded; thus the bond futures market has contracts for March, June, September and December. (4) *Location*: the lot must be delivered in specified places (e.g. warehouses or banks) in one or more specified cities. (5) *Identity of contractors*: after the initial contract is established, the buyer and seller normally have no further dealings with each other, thus eliminating credit risk. The execution is guaranteed by a clearing house, which acts as seller to all buyers and as buyer to all sellers. The clearing house can offer this guarantee by virtue of the security deposits, known as 'margin', it collects from its members.

The immediate purpose of this standardization is to minimize transaction costs and thereby to endow the futures contract with the ready negotiability that forward contracts, heterogeneous as they are, normally lack. Futures contracts are intended to be traded by 'open outcry' on the floor of an organized exchange. Such exchanges are found in a number of commercial centres, especially in Chicago, New York and London.

The overall market for a commodity or financial instrument can be divided into the futures

market, which is centralized and trades only standardized contracts, and the cash market, which is dispersed and deals in actual parcels of the commodity or instrument. The cash market can be further divided into the spot market and the forward market.

Traders may have long or short positions in any or all of these three markets; thus a merchant who holds a physical inventory is considered to be long in the spot market. A trader whose net position in the case market is offset by his position in the futures market is called a *hedger*; more particularly he is a 'short hedger' if he is long in the cash market and short in the futures market, and a 'long hedger' if these positions are reversed. Traders who are net long or net short in the overall market (and hence in at least one of its submarkets) are known as *speculators*. In the futures market there also 'spreaders' or 'straddlers', whose long position in one or more futures contracts exactly matches their short position in other futures contracts.

In both the futures and the forward markets the net position of all traders combined must be zero, since there is a sale for every purchase. This is not true in the spot market, where the aggregate net position is positive to the extent of the existing inventories. The total of all long (or short) positions in the futures market is called the 'open interest'.

The prices prevailing in the cash and futures markets at any time are not necessarily equal. However, there are two main links between these markets; one is provided by the delivery mechanism and the other by hedging. As to delivery, when a futures contract reaches maturity (as the May contract does in the month of May) the remaining shorts have to deliver what they have sold, and the remaining longs have to accept and pay for what they have bought. Clearly the shorts will not deliver anything that could be sold at a higher price in the spot market, nor will the longs take delivery of anything that they could buy more cheaply elsewhere. At delivery time, therefore, the futures price must be equal to the spot price of the items that are actually delivered. Since this ultimate equality is widely anticipated, it will also influence futures and spot prices prior to delivery time.

Hedging also serves to relate futures prices and spot prices. As Working (1953) pointed out, it is essentially a form of arbitrage between the two markets. If a futures price is high compared to a spot price, hedgers will buy in the spot market and sell futures. They can do so without risk if the futures price exceeds the spot price by more than the *carrying charge*, which is the cost of holding physical inventories between the present and the maturity of the futures contract. The futures price therefore cannot exceed the current spot price by more than the prevailing carrying charge.

It does not follow, however, that a futures price must always exceed the spot price by the relevant carrying charge. Positive inventories may be held even if the spot price is above the futures price. This is because inventories have what Kaldor (1939) called a 'convenience yield', derived from their availability when buyers need them. The profits of merchants, in fact, depend in large part on their ability to assess and realize the convenience yield. Its size depends primarily on the size of total inventories; if they are small, the marginal convenience yield will be high, but if they are large, it may be zero. Working (1953) described the relationship between the size of inventories and the return of them as the *supply curve of storage*.

The view of hedging expressed above is not necessarily inconsistent with the older interpretation of hedging as an effort to shift the price risk inherent in holding inventories to those (namely the speculators) willing to assume this risk in the hope of profiting from favourable price movements. It should be noted, however, that hedging need not reduce the total risk to which a hedger is exposed. Bankers are generally willing to finance a larger proportion of the value of hedged inventories than of unhedged inventories. By hedging, consequently, a merchant can support a larger inventory with his own capital, thereby giving more scope to the exercise of his merchandising skills. The connection between hedging and risk aversion is not as clear-cut as the older view would suggest.

Regardless of the economic interpretation of hedging, its existence has another important implication discovered by Keynes (1923, 1930) and elaborated by Hicks (1939) and Houthakker

(1968). If merchants can increase their profits by hedging, they must be willing to pay a *risk premium* for the opportunity to do so. It is conceivable that short hedging (defined above) exactly offsets long hedging, in which case any premiums paid by hedgers would cancel out. There is considerable evidence, however, that in most markets short hedging exceeds long hedging at most times. The basic reason for this asymmetry is that, as pointed out earlier, the net position in the spot market (and hence in the overall market) is positive. In seasonal commodities an excess of long hedging over short hedging is usually found only towards the end of the crop year, when inventories are small.

Now if the hedgers are net short in futures, the speculators in futures must be net long. Keynes and his followers argued that speculators will only be net long if they expect futures prices to rise. At any particular moment the speculators may of course be wrong, but on the average they are right, and each futures price will tend to rise until, at the maturity of the contract, it equals the relevant spot price. The speculators' gain is the hedgers' loss; thus the speculators receive a risk premium proportionate to the amount of hedging they make possible. This risk premium is implicit in the hedgers' willingness to sell futures contracts that have a tendency to appreciate.

This, in brief, is Keynes' theory of *normal backwardation*. ('Backwardation' designates a situation where the futures price is below the spot price; strictly speaking the term 'normal backwardation' applies only to the nonseasonal markets that Keynes had in mind, but the fundamental idea carries over to markets with seasonality.) The theory anticipated the positive relation between risk and return that is the main result of the Capital Asset Pricing Model developed in the 1960s. Consistency with CAPM also requires, however, that the risk of buying futures cannot be eliminated by diversification, and that has not yet been demonstrated. The theory of normal backwardation can also be summarized as saying that futures prices, when viewed as predictors of the spot price in the future, have a downward bias.

The empirical validity of the theory of normal backwardation remains in dispute. Favourable evidence has been presented by Houthakker (1957, 1961, 1968), Cootner (1960) and Bodie and Rozansky (1980). For adverse evidence see Telser (1958, 1981), Gray (1961), Rockwell (1967) and Dusak (1973). According to the latter group of authors, futures prices are unbiased predictors of spot prices, and no risk premium is paid. The most telling argument of the critics of normal backwardation is that as a body, small speculators appear to lose money rather consistently.

If true, the theory of normal backwardation would also shed light on an observation made earlier, namely the fairly limited scope of futures trading. To be viable, the theory implies, a futures market has to be nourished by the risk premium transferred from the hedgers to the speculators; in its absence the latter would be gradually driven out by the transaction costs they incur. The futures contract must therefore be primarily designed to attract hedging.

It is not a simple matter to design futures contracts that will attract enough hedging to ensure their continued viability. Hedgers need a high correlation between the futures prices and the particular spot prices in which they are interested; consequently the contract should be neither too broad (i.e. include too many deliverable grades) nor too narrow. There must also be enough variability in prices to make hedging and speculation worthwhile.

This is why futures trading was for many years confined to grains, oilseeds, sugar, cotton, non-ferrous metals and a few other staples that can be easily graded and have volatile prices. There is no futures trading in such important commodities as steel, paper and synthetic fibres. In the 1970s, when exchange rates and interest rates became more variable, futures trading was successfully introduced in various financial instruments – first in foreign exchange, then in government securities and similar claims, and most recently in indexes of share prices. Financial futures now account for most of the activity in futures markets. The most important recent addition in the non-financial sector has been futures trading in crude oil and some of its derivatives.

Despite the controversy over normal backwardation it is widely agreed that one of the

economic functions of futures trading is risk transfer. Another such function is sometimes called *price discovery*. It consists in the establishment of a competitive reference price for a commodity or financial instrument. Since the cash market is typically heterogeneous, it is convenient to have a single price from which spot and forward prices can be derived as differences. Thus the forward price for a specific transaction may be quoted as a number of cents over or under the May futures price.

Futures trading also facilitates the *allocation of production and consumption over time*, particularly by providing market guidance in the holding of inventories through the supply curve of storage (see above). More generally futures prices provide information relevant to the planning of production and consumption; if the futures prices for distant deliveries are well below those for early delivery, for instance, postponing consumption is more attractive.

The economic functions of futures markets will be performed most effectively when they are highly competitive. If one or more traders are large enough to assert their market power, futures prices (and quite possibly cash prices) may not reflect the underlying supply and demand conditions. The prevention of such distortions – particularly of 'corners', where one or more longs manipulate both the cash and the futures market – is a major concern of futures exchanges and their regulators. In the United States the Commodity Futures Trading Commission supervises the markets with a view to preventing and penalizing these and other abuses, though it has not always succeeded. In Britain the Bank of England has somewhat similar responsibilities.

See Also

- ▶ Backwardation
- ▶ Futures Markets, Hedging and Speculation
- ▶ Hedging

Bibliography

Bodie, Z., and V.J. Rozansky. 1980. Risk and return in commodity futures. *Financial Analysts' Journal* 36(27–31): 33–39.

Cootner, P. 1960. Returns to speculators: Telser *vs.* Keynes. *Journal of Political Economy* 68: 396–418 (with reply by Telser and rejoinder by Cootner).

Dusak, K. 1973. Futures trading and investor returns: An investigation of commodity market risk premiums. *Journal of Political Economy* 81(6): 1387–1406.

Gray, R. 1961. The search for a risk premium. *Journal of Political Economy* 69: 250–260.

Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.

Houthakker, H.S. 1957. Can speculators forecast prices? *Review of Economics and Statistics* 39: 143–152.

Houthakker, H.S. 1961. Systematic and random elements in shorterm price movements. *American Economic Review, Papers and Proceedings* 51: 164–172.

Houthakker, H.S. 1968. Normal backwardation. In *Value, capital and growth*, ed. J.N. Wolfe. Edinburgh: Edinburgh University Press.

Kaldor, N. 1939. Speculation and economic stability. *Review of Economic Studies* 7: 1–27.

Keynes, J.M. 1923. Some aspects of commodity markets. *Manchester Guardian Commercial, Reconstruction Supplement* 29, March. Reprinted in *The collected writings of John Maynard Keynes*, Vol. VII, London: Macmillan, 1973.

Keynes, J.M. 1930. *A treatise on money*, vol. II. London: Macmillan.

Rockwell, C.S. 1967. Normal backwardation, forecasting and the return to commodity futures traders. *Food Research Institute Studies* 7: 107–130.

Telser, L.G. 1958. Futures trading and the storage of cotton and wheat. *Journal of Political Economy* 66: 233–255.

Telser, L.G. 1981. Why there are organized futures markets. *Journal of Law and Economics* 24(1): 1–22.

Working, H. 1953. Hedging reconsidered. *Journal of Farm Economics* 35: 544–561.

Fuzzy Sets

Claude Ponsard

The scope of fuzzy economics is to bring into play a new body of concepts in which imprecision (or fuzziness) is accepted as a matter of science. Accurate mathematical methods are used; they are based on the concept of *fuzzy set*. Intuitively, a fuzzy set is compounded of elements which appertain to it *more or less*. The transition from membership to non-membership is soft rather than crisp, as in the case of an ordinary set. In the same manner, *fuzzy logic* handles imprecise

truths, and fuzzy connectives and rules of inference, contrary to classical two-valued logic.

The theory of fuzzy sets was initiated by Zadeh (1965). Since then the literature has been plentiful but scattered. Periodically, some handbooks have gathered important results (Kaufmann 1975; Dubois and Prade 1980; Zimmermann 1985).

The word *fuzzy set* is a misuse of language. More exactly, the proper term is *fuzzy subset* because the reference set is not fuzzy. In what follows ordinary (non-fuzzy) concepts are in bold italic, whereas fuzzy concepts are not. For example, $X \subset \boldsymbol{E}$ is read: X is a fuzzy subset of the ordinary reference set $\boldsymbol{E}$.

Let $\boldsymbol{E} = \{x\}$ be a non-empty, finite or not, set and $\boldsymbol{M}$ a preordered set, with Card $\boldsymbol{M} \geq 2$. Let $\boldsymbol{M}^{\boldsymbol{E}}$ be the set of the mappings from $\boldsymbol{E}$ into $\boldsymbol{M}$. By definition a fuzzy subset X of the reference set $\boldsymbol{E}$ is an element of $\boldsymbol{M}^{\boldsymbol{E}}$ such that $X = \{x, \mu_X; \forall x \in \boldsymbol{E}: \mu_X(x) \in \boldsymbol{M}\}$, where μ_X is a mapping from $\boldsymbol{E}$ into $\boldsymbol{M}$. The mapping $\mu_X(x)$ is called the membership function of x to X and expresses the degree of membership of the element x of $\boldsymbol{E}$ to the fuzzy subset X of $\boldsymbol{E}$.

Many particular fuzzy subsets theories can be stated according to the characterization of the membership set $\boldsymbol{M}$. First, $\boldsymbol{M}$ is a non-numerical set; its elements are linguistic variables which are applied to approximate reasoning (Zadeh 1975). Second, $\boldsymbol{M}$ is a set of ordinary numbers; then different fuzzy subsets can be defined according to the structure of each particular set of numbers which is chosen as membership set. For elaborating theoretical properties and empirical applications, it is convenient to make a distinction depending on whether $\boldsymbol{M}$ is a lattice or a lattice of intervals. In numerous theoretical statements and in the quasi-totality of applications, $\boldsymbol{M} = [0, 1]$. This characterization was initially proposed by Zadeh (1965). In the most general case, $\boldsymbol{M}$ can be any lattice (Goguen 1967); fuzzy subsets having more or less general properties are defined, according to the properties of lattices: distributive, complemented, boolean lattices, etc. If $\boldsymbol{M}$ is a lattice of intervals, denoted by $[a_i, a_j] \subseteq [0, 1]$, then still more general fuzzy subsets can be defined. Sambuc (1975) has initially stated the theory, named phi-fuzzy subsets theory. The value of the membership function, denoted by $\Phi_x(x) = [a_i, a_j]$, is equal to the whole interval $(a_i, a_j]$, not to a number included into the interval. Of course, other particular specifications using a set of ordinary numbers as membership set can be stated.

Now $\boldsymbol{M}$ can be a set of fuzzy numbers, whose theory was initiated by Dubois and Prade (1980). A fuzzy number expresses that the value of a variable is not exactly equal to a precise number; the exact value is more or less credible. Consider a fuzzy membership function, denoted by μ_n, from $\mathbb{R}$. into [0, 1] and such that $\forall x \in \mathbb{R} : \mu_x(x) \in [0, 1]$. Thus n is a fuzzy subset of R. If the two following conditions are fulfilled: μ_n has the normality property and is quasi-concave, then the associated fuzzy subset n is called a fuzzy number.

All these specifications must be carefully distinguished because most of the properties of fuzzy subsets are induced by that of the membership set $\boldsymbol{M}$. Furthermore, in applications, if $\boldsymbol{M}$ is a set of ordinary numbers, the fuzziness which is associated with a datum is expressed in an exact manner, whereas it is expressed in a fuzzy manner when $\boldsymbol{M}$ is a set of fuzzy numbers (Ponsard 1985b).

Of course, if $\boldsymbol{M} = \{0, 1\}$, ordinary set theory is found again, as a particular case.

The axiomatic framework of fuzzy subsets theory includes that of the theory of measurable sets. A fuzzy measure is defined on a fuzzy σ-algebra over the reference set. A fuzzy σ-algebra differs from a σ-algebra owing to the fact that it does not have the property of complementation. A fuzzy measure on a fuzzy σ-algebra is a mapping with co-domain a preordered and bounded set satisfying some axioms which are less restrictive than the conditions required for an ordinary measure. In particular, a fuzzy measure need not be additive.

So, a careful distinction must be made between the theory of fuzzy subsets and the theory of probability. A probability measure is a mapping from a σ-algebra (with the complementation property) over the reference set into $\mathbb{R}^+$ such that the additive property, among all the axioms, is necessarily verified. Concepts of fuzziness and risk being distinguished, a theory of fuzzy random sets which handles the probabilities of fuzzy events can be stated (Zadeh 1968).

Finally, in the same manner, the relation between the concepts of fuzziness and uncertainty have to be settled (Zadeh 1978). A distribution of *possibilities* is a function, denoted by φ from E into [0, 1] such that

$$\operatorname*{Sup}_{x \in \boldsymbol{E}} \varphi(x) = 1.$$

Possibilities are not additive, contrary to probabilities. Clearly, the theory of risk (or probability) formulates what *must* occur, whereas the theory of uncertainty (or possibility) expresses what *may* happen.

In economics, fuzzy analysis was initiated by Ponsard (1975). Then the Institute of Economic Mathematics (University of Dijon, France) devoted a programme to the field in the framework of spatial economic analysis. Ponsard (1983) specified the place of fuzzy space analysis in the context of modern spatial economic theory.

Many types of fuzzy economic spaces were studied by several contributors: attraction zones for sale-points, areas of fuzzy spatial interactions, fuzzy regional dynamic systems, fuzzy interregional relations, fuzzy urban spaces, mental maps, etc. Indeed, the description of economic spaces has now at its disposal pertinent and sophisticated mathematical tools. For example, in regional analysis, Tranqui (1978) states an automatic classification method which integrates fuzzy data on the observed territories and applies it to the French economy. Then Ponsard and Tranqui (1985) apply the same method to the European economy. More or less fine subdivisions result as a function of the more or less strictness of the chosen degree of similarity and described regions are separated or overlapped. From a complementary point of view, economic regions are analysed as a central places system, where agglomerations are linked together by flows which generate a set of numerous interrelations. The influences exerted by each agglomeration on the others are diffuse and vague by nature. So, the use of several indicators allows us to surround the minimal and maximal bounds of the magnitude of each influence relation in a realistic manner. Ponsard (1977) builds up a phi-fuzzy network such that the arcs which join any pair of agglomerations are valued by an interval which expresses the margin of fuzziness in a given influence relation. In this framework, the fuzzy hierarchical structure of a central places system is revealed.

Besides fuzzy spaces, the analysis of fuzzy spatial behaviours is an important and complementary field whose scope is to state the microeconomic foundations of macroeconomic spaces and the conditions for partial and general equilibria.

In the present state of the art, the locations of economic agents are given, so that partial equilibria are analysed in terms of produced and exchanged quantities of goods, and the general equilibrium in terms of quantities and prices. Three stages have to be distinguished.

First, the economic agent does not generally manifest a perfect aptitude to discriminate clearly, among alternatives between those he prefers and those he does not prefer. It follows that his behaviour does not obey a binary logic of the type preference-non-preference, but a fuzzy logic (Ponsard 1981a; 1985a). Let $\boldsymbol{E} = \{x_i\}$ be a set of a priori possible alternatives. The behaviour of the economic agent is characterized by a structure $(E, \mathfrak{R})$ where $\mathfrak{R}$ is a fuzzy binary relation between the elements of $\boldsymbol{E}^2$. It is such that:

$$x_i \mathfrak{R} x_j = \left\{ \left(x_i, x_j\right), \mu_{\mathfrak{R}}; \forall x_i \in \boldsymbol{E}, \forall x_i \in \boldsymbol{E} : \mu_{\mathfrak{R}}\left(x_i, x_j\right) \in \boldsymbol{M} \right\}$$

where $\boldsymbol{M}$ is a preordered and bounded membership set and $\mu_{\mathfrak{R}}(x_i, x_j)$ expresses the degree of fuzziness which characterizes the correspondence between two given alternatives. The structure $(\boldsymbol{E}, \mathfrak{R})$ has many interesting properties: a strong degree of preference for x_i with respect to x_j can be distinguished from a weak degree of preference for x_j with respect to x_i, fuzzy reflexivity property, Max–Min transitivity property (whose definition is weaker than the classical one), totality property (so that non-comparability does not raise specific problems). Finally, a fuzzy total preorder on $\boldsymbol{E}$ is obtained; the classes of indifference are antisymmetrical and, as such, they form between themselves a fuzzy order relation. Then, under some

F

conditions which assure the existence of a fuzzy topological totally preordered space, a fuzzy continuous utility function, denoted by μ_u, is stated. Now, $\boldsymbol{M} = [0, 1]$ in order to a numerical representation of preference be determined. The utility function is such that, $\forall x_i \in \boldsymbol{E}$, $\mu_u(x_i) \in [0, 1]$.

Thus the theory of fuzzy spatial preference and utility is neither ordinal nor cardinal. The functions taking their values in any set $\boldsymbol{M}$ or in the interval [0, 1] are fuzzy measures so that ordinal and cardinal theories are particular cases of this *valuation* theory, *valuation* being taken to mean fuzzy measure in short.

Second, the models of fuzzy spatial equilibria of consumer and producer are based on specifications which are peculiar to their respective fields (Ponsard 1981b; 1982a). They are particular cases of the economic calculation of optimizing a fuzzy objective function under an elastic resource limitation constraint. Again let $\boldsymbol{E}$ be a set of alternatives. A fuzzy decision, denoted by D, in $\boldsymbol{E}$ is by definition the intersection of the fuzzy subset F, $F \subset \boldsymbol{E}$, describing the aimed objective, and the fuzzy subset C, $C \subset \boldsymbol{E}$, describing the constraint. So $D = F \cap C$ with a membership function, denoted by μ_D, such that,

$$\forall x \in \boldsymbol{E}, \mu_D(x) = \mu_F(x) \wedge \mu_C(x), \quad \text{with } \mu_D(x) \simeq 1 \text{ iff } x$$

is good for F and C and $\mu_D(x) \simeq 0$ iff x is bad for F or C. In fuzzy algebra, the intersection operation makes use of the *M in* operator (denoted by Λ). Then an optimal decision is such that:

$$\operatorname*{Sup}_{x \in \boldsymbol{E}} \mu_D(x) = \operatorname*{Sup}_{x \in \boldsymbol{E}} [\mu_F(x) \wedge \mu_C(x)].$$

This formulation calls on an important remark in the framework of spatial partial equilibria theories: objective and constraint are two fuzzy subsets of the same reference set and have the same role in decision making; their relations are symmetrical since the intersection operation is commutative.

Tanaka et al. (1974) have proved that the solution for the problem of finding the best possible decision is to select an element x in E such that:

$$\operatorname*{Sup}_{x \in \boldsymbol{E}} \mu_D(x) = \operatorname*{Sup}_{x \in \boldsymbol{A}} \mu_F(x),$$

with $\boldsymbol{A} \subset \boldsymbol{E}$ and $\boldsymbol{A} = \{x; x \in \boldsymbol{E} : \mu_C(x) \geq \mu_F(x)\}$. In clear language, $\boldsymbol{A}$ is a non-fuzzy subset of $\boldsymbol{E}$ such that the value of the constraint membership function is at least equal to the value of the objective membership function. The conditions for the function

$$\operatorname*{Sup}_{x \in \boldsymbol{A}} \mu_F(x)$$

to be continuous are only mildly restrictive. Among them, the condition that the fuzzy subset which describes the objective be strictly convex (in the weaker sense of convexity in fuzzy analysis). Mathematically, it would be indifferent to place the strict convexity condition on the constraint rather than the objective, since they have the same part in the decision making. In economic analysis, it is accurate to place it on the objective. Indeed, in the consumer and producer spatial equilibria theories, it guarantees the continuity property of the fuzzy objective functions. Moreover, in producer equilibrium theory, the awkward situation in which returns are increasing does not pose a specific problem since the strict convexity condition is not placed on the technological constraint. Moreover, the solution is generally not unique, which is an expected result in a fuzzy context. Finally, in the particular case where the objective is precise and the constraint alone is fuzzy, then the fuzzy economic calculation can be solved by a different and much simpler method (Ponsard 1982b).

In the third stage, a theory of spatial general equilibrium with fuzzy behaviours is stated (Ponsard 1984). Excess demand, denoted by e, is dependent on a spatial delivered price system, denoted by p (a price vector). So, an excess demand fuzzy point-to-set mapping denoted by ϕ is defined from $(\boldsymbol{P} \times \boldsymbol{E})$ to where $\widehat{\boldsymbol{P}}$ designates the set of standard prices and the fuzzy power-set of $\left(\widehat{\boldsymbol{P}} \times \boldsymbol{E}\right)$. At the equilibrium, the condition that $e \leq \mathrm{O}$ has to be verified. The conditions which ought to be fulfilled by $e(p)$ in order for p to be such that $e(p) \leq \mathrm{O}$ exists, must be stated.

The analysis is based on Butnariu's theorems (1982) which extend Brouwer's and Kakutani's theorems to fuzzy functions and fuzzy point-to-set mappings respectively. Economic results are the generalization of Walras's Law to an economic space where behaviours are soft, and the formulation of the following theorem: if the excess demand fuzzy point-to-set mapping is closed and has images which are non-empty, normal and convex, and verifies the generalized Walras's Law, then a competitive equilibrium exists, i.e. there exist a price vector $p^* \in \boldsymbol{P}$ and an excess demand vector $e^* \in e(p^*)$ such that $e^* \leq 0$. This theorem is a generalization of a famous result of Debreu (1959) to the case of a spatial economy characterized by fuzzy behaviours of agents. It is true whatever the distribution of locations. Finally, the concept of fuzzy expected utility which brings into play fuzzy random sets and possibility theory is stated by Mathieu-Nicot (1985).

In fact, the chief difficulty is to determine the membership function and the fuzzy measure for the fuzzy subsets of a referential. Of course, there exist no general and unique method. A solution must be found in every case. However this difficulty is not peculiar; in the same manner, the determination of a distribution of probability in stochastic models is often hard.

Finally, it is easy to look forward to further research not only in the field of spatial analysis, but also in general economic theory.

Bibliography

Butnariu, D. 1982. Fixed points for fuzzy mappings. *Fuzzy Sets and Systems* 7(2): 191–207.

Debreu, G. 1959. *Theory of value: An axiomatic analysis of economic equilibrium*, Cowles Foundation Monograph, vol. 17. New York: John Wiley & Sons.

Dubois, D., and H. Prade. 1980. *Fuzzy sets and systems: Theory and applications*. New York: Academic Press.

Goguen, J.A. 1967. L-fuzzy sets. *Journal of Mathematical Analysis and Applications* 18: 145–174.

Kaufmann, A. 1975. *Introduction to the theory of fuzzy subsets*, Fundamental Theoretical Elements, vol. 1. New York: Academic Press. (trans. of French edn of 1973).

Mathieu-Nicot, B. 1985. *Espérance mathématique de l'utilité floue*, Coll. IME, vol. 29. Dijon: Librairie de l'Universite.

Ponsard, C. 1975. L'imprécision et son traitement en analyse économique. *Revue d'Economie Politique* 1: 17–37.

Ponsard, C. 1977. Hiérarchie des places centrales et graphes phi-flous. *Environment and Planning A* 9: 1233–1252.

Ponsard, C. 1981a. An application of fuzzy subsets theory to the analysis of the consumer's spatial preferences. *Fuzzy Sets and Systems* 5(3): 235–244.

Ponsard, C. 1981b. L'équilibre spatial du consommateur dans un contexte imprécis. *Sistemi Urbani* 3: 107–133.

Ponsard, C. 1982a. Producer's spatial equilibrium with a fuzzy constraint. *European Journal of Operational Research* 10: 302–313.

Ponsard, C. 1982b. Partial spatial equilibria with fuzzy constraints. *Journal of Regional Science* 22: 159–175.

Ponsard, C. 1983. *History of spatial economic theory*, Texts and Monographs in Economics and Mathematical Systems. Berlin: Springer-Verlag.

Ponsard, C. 1984. *A theory of spatial general equilibrium in a fuzzy economy.* Working Paper No. 65, IME. Revised version in *Fuzzy economics and spatial analysis*, ed. C. Ponsard and B. Fustier, Coll. IME 32. Dijon: Librairie de l'Université, 1986.

Ponsard, C. 1985a. Fuzzy sets in economics: Foundation of soft decision theory. In *Management decision support systems using fuzzy sets and possibility theory*, Coll. Interdisciplinary Systems Research, vol. 83, ed. J. Kacprzyk and R.R. Yager, 25–37. Cologne: Verlag TUV Rheinland.

Ponsard, C. 1985b. Fuzzy data analysis in a spatial context. In *Measuring the unmeasurable*, Series D, no. 22, NATO ASI Series, ed. P. Nijkamp, H. Leitner, and N. Wrigley, 487–508. Dordrecht: Martinus Nijhoff.

Ponsard, C., and P. Tranqui. 1985. Fuzzy economic regions in Europe. *Environment and Planning, Series A* 17: 873–887.

Sambuc, R. 1975. *Fonctions phi-floues. Application á l'aide au diagnostic en pathologie thyroïdienne*. PhD thesis, Université de Marseille.

Tanaka, H., T. Okuda, and K. Asai. 1974. On fuzzy mathematical programming. *Journal of Cybernetics* 3: 37–46.

Tranqui, P. 1978. *Les régions économiques floues: Application au cas de la France*, Coll. IME, vol. 16. Dijon: Librairie de l'Université.

Zadeh, L.A. 1965. Fuzzy sets. *Information and Control* 8: 338–353.

Zadeh, L.A. 1968. Probability measures of fuzzy events. *Journal of Mathematical Analysis and Applications* 23: 421–427.

Zadeh, L.A. 1975. The concept of a linguistic variable and its application to approximate reasoning. *Information Sciences*, Part 1: 8, 199–249; Part 2: 8, 301–57; Part 3: 9, 43–80.

Zadeh, L.A. 1978. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems* 1(1): 3–28.

Zimmermann, H.J. 1985. *Fuzzy set theory and its applications*. Dordrecht: Kluwer-Nijhoff.