G

Gains from Trade

Murray C. Kemp

Questions relating to the gainfulness or otherwise of international trade and investment have always interested economists, from Adam Smith to the present day. We now have at our disposal a very large arsenal of propositions concerning the trading gains of single countries and of groups of countries under alternative institutional arrangements. However, most of these propositions relate to the limiting case of small countries. For example, much ingenuity has been expended in tracking the welfare implications of autonomous changes in the world prices faced by a small country or in the vector of tariffs imposed by such a country. Evidently the fruits of such investigations are of only modest general interest. Here we concentrate on two propositions which are valid for economies of any size and which are of considerable historical and intellectual interest. For an accurate summary of small-country results, and for the relevant references to the literature, see Woodland (1982, chs 9 and 11).

The Benefits of Free and Competitive Trade

We begin with the oldest and best-known of all propositions in the literature concerning the gains from trade, indeed in the history of economic thought.

Proposition 1

If an initially autarkic or non-trading country s is exposed to free commodity trade with one or more other countries, either in the whole set of producible goods or in some subset, and if preferences, technologies and endowments are restricted in the manner of Arrow and Debreu (1954) and if markets are complete, then there is a competitive world trading equilibrium (possibly with lump sum transfers within s) such that no individual in s is worse off than in autarky.

Proposition 1 is widely accepted. However, it is not immediately plausible. Of course the opening of trade between countries enlarges the set of feasible worldwide consumption vectors. It then follows from the Second Theorem of Welfare Economics that there exists a competitive world equilibrium possibly with lumpsum transfers, such that no individual is worse off than under universal autarky. It might be thought therefore that there is nothing to understand and nothing to prove, that Proposition 1 is embedded in a standard theorem of welfare economics. However, in the statement of transfers theorem there are no restrictions on the scope of transfers whereas in the statement of Proposition 1 transfers are required to balance within each country. Thus there is indeed something to prove.

Nor is the proof easy. Indeed it was not until 1972, nearly two hundred years after the *Wealth of Nations*, that formal and general statements and proofs became available (see Grandmont and

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

McFadden 1972; Kemp and Wan 1972). One reason for the long lag between conjecture and proof is, undoubtedly, the technical difficulty of establishing the existence of a lump-sum compensated world equilibrium; the appropriate tools for such a demonstration became known to economists only after World War II.

It has been noted that Proposition 1 rests on assumptions of Arrow–Debreu (1954) type. In particular, the number of goods is required to be finite and the set of markets complete. Without both of those assumptions there is no assurance that free trade is gainful to all participating countries. Kemp and Long (1979) have shown that in an infinite-horizon model with overlapping finite generations, and therefore with an infinity of dated goods, trade can be unambiguously harmful to one of the trading partners, even though all countries are competitive and free of conventional distortions, externalities, non-convexities and learning processes. Of course, Malinvaud (1953, 1962) had shown long ago that closed economies of the type studied by Kemp and Long can be inefficient; and it is not surprising perhaps that trade between inefficient economies is not always mutually gainful. Similarly, Newbery and Stiglitz (1981, ch. 23) have shown that if there is an incomplete set of markets in each trading country, so that the several autarkic equilibria are inefficient, then the opening of trade can leave every individual worse off.

Moreover, it is essential to the conclusions of the proposition that compensation be lumpsum. In their recent book, Dixit and Norman (1980) appear to suggest that if trade is strictly gainful with lumpsum compensation then it is strictly gainful with compensation effected by carefully chosen (non-lumpsum) taxes on goods. The suggestion is an interesting one for, if valid, it would imply that any internal misallocation generated by the (carefully chosen) commodity taxes is always more than offset by the possibility of trading at world prices. However, it has been shown by counter example that the suggestion is ill-founded, that nonlumpsum compensation is an adequate substitute for lumpsum compensation only in special cases; see Kemp and Wan (1986a).

Proposition 1 affirms that, for each participating country, free trade is preferable to autarky. It does not state that, for each country, free trade is preferable to all other kinds of trade. Indeed it was recognized quite early, by Sir Robert Torrens (1821, 1844) and John Stuart Mill (1844), that a large trading country, with market power, can improve its position by manipulating its trade with the aid of taxes and subsidies on its exports and imports; indeed, by offering all-or-nothing contracts a large country can do even better than indicated by Torrens and Mill.

The Welfare Economics of Customs Unions

The interest of economists in customs unions goes back at least to the Prussian Zollverein of 1819–31. For the most part, however, that interest has focused on the trade-distorting effects of unions rather than their welfare-distorting effects. Indeed, it was not until quite recently that a welfare proposition of any generality was established. The following proposition was first stated by Kemp (1964) and later proved under Arrow–Debreu assumptions by Kemp and Wan (1976, 1986b).

Proposition 2

Consider any competitive world trading equilibrium with any number of countries and any finite number of commodities, and with no restrictions on the tariffs and other commodity taxes of individual countries. Let any subset of the countries form a customs union. Then there exists a common tariff vector and a system of lumpsum compensatory payments, involving only members of the union, such that there is an associated competitive equilibrium in which each individual, whether a member of the union or not, is not worse off than before the formation of the union. Proposition 2 has been extended by Grinols (1981) who displayed a particular scheme of compensation based on observable features of the pre-union equilibrium only.

Proposition 2 shows that there is an incentive for trading countries to move towards worldwide free trade, the ultimate customs union. That we do not observe a free-trading world, or even an unmistakable drift to free trade, can be traced to game-theoretical conflicts about the choice of partners, the division of the gains and the enforcement of agreements; to the non-economic objectives of nations; and to the unrealism of some of the Arrow–Debreu assumptions, notably the assumptions that there are no externalities and that production sets and preferences are convex.

See Also

- ▶ Foreign Trade
- ▶ International Trade

Bibliography

Arrow, K.J., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Dixit, A.K., and V. Norman. 1980. *Theory of international trade*. Welwyn/Herts: J. Nisbet/Cambridge University Press.

Grandmont, J.M., and D. McFadden. 1972. A technical note on classical gains from trade. *Journal of International Economics* 2(2): 109–125.

Grinols, E.L. 1981. An extension of the Kemp–Wan theorem on the formation of customs unions. *Journal of International Economics* 11(2): 259–266.

Kemp, M.C. 1964. *The pure theory of international trade*. Englewood Cliffs: Prentice-Hall.

Kemp, M.C., and N.V. Long. 1979. The under-exploitation of natural resources: A model with overlapping generations. *Economic Record* 55: 214–221.

Kemp, M.C., and H.Y. Wan Jr. 1972. The gains from free trade. *International Economic Review* 13(3): 509–522.

Kemp, M.C., and H.Y. Wan Jr. 1976. An elementary proposition concerning the formation of customs unions. *Journal of International Economics* 6(1): 95–97.

Kemp, M.C., and H.Y. Wan Jr. 1986a. Gains from trade with and without lumpsum compensation. *Journal of International Economics* 21(1–2): 99–110.

Kemp, M.C., and H.Y. Wan Jr. 1986b. The comparison of second-best equilibria: The case of customs unions. In *Zeitschrift fur Nationalökonomie*. Vienna: Springer.

Malinvaud, E. 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21: 233–268.

Malinvaud, E. 1962. Efficient capital accumulation: A corrigendum. *Econometrica* 30(3): 570–573.

Mill, J.S. 1844. *Essays on some unsettled questions of political economy*. London: John W. Parker.

Newbery, D.M.G., and J.E. Stiglitz. 1981. *The theory of commodity price stabilization: A study in the economics of risk*. Oxford: Oxford University Press.

Torrens, R. 1821. *An essay on the production of wealth*. London: Longman, Hurst, Rees, Orme and Brown.

Torrens, R. 1844. *The budget. On commercial and colonial policy*. London: Smith, Elder & Co.

Woodland, A.D. 1982. *International trade and resource allocation*. Amsterdam: North-Holland.

Gaitskell, Hugh Todd Naylor (1906–1963)

M. Anyadike-Danes

Hugh Gaitskell was born in 1906 in London and educated at Winchester and New College, Oxford. The General Strike, which occurred mid-way through his undergraduate studies, led to Gaitskell's first active involvement in politics when he assisted local supporters of the Trade Union Council: this experience, and the aftermath of the Strike, began his life-long commitment to the labour movement. Having graduated in 1927 with first class honours in 'Modern Greats' (Politics, Philosophy and Economics) his first job was as Workers' Educational Association lecturer at University College, Nottingham, but after only a year's teaching there he was offered, and accepted, a post as lecturer in economics in the Department of Political Economy at University College, London. The move south did not, however, stem from any desire to pursue a more conventional university career, as he wrote to his mother from Nottingham in the spring of 1928:

> I shall probably not become Academic for (a) I dislike the academics and their attitude and their bourgeoisieness (b) I am likely to continue my association with the Labour movement. I have seen enough of Working Class conditions, industrial war and Class war here to make it probable that on and off through my life . . . I shall be taking part in the Working Class movement. (quoted by Williams 1982, p. 36)

Most of Gaitskell's research and writing on economic theory and policy was done during the next 11 years, spent at University College. His academic output was not prolific and most of it was concerned with the 'Austrian' approach to economic theory: he published two highly regarded papers on the period of production (in German); contributed to the translation of Haberler's *Theory of International Trade* (1935); and began, but never completed, the translation of some of Bohm-Bawerk's writings on capital theory. He played a very active role during this period in the formulation, discussion and dissemination of Labour Party economic policy. In particular, he was a leading member of the New Fabian Research Bureau whose activities in the 1930s grouped together a wide circle of the younger socialist-inclined economists. An indication of Gaitskell's views of the appropriate policy response to the problem of mass unemployment is provided in his essay 'Financial Policy in the Transition Period' which appeared in 1935. Most of the paper was concerned with the policies which an incoming Labour government might adopt to counter the 'financial panic' which it was widely believed would accompany their election, but it also contained some more general remarks on the nature of the 'expansionist programme' which a Labour government should pursue:

> The efficacy of monetary policy as a method of curing industrial depression is still a matter of controversy. But that at certain times the banking system as a whole has the power to stimulate industrial expansion can scarcely be questioned, ... There is no doubt, for example, that the very moderate measure of recovery achieved by this country is due in the main to the abandonment of the gold standard and the subsequent policy of the Bank of England. This policy has been of the 'orthodox' character of simply creating and maintaining low rates of interest through the instruments of bank rate and open-market policy. . . . But although a low long-term rate is certainly essential . . . its action is always very slow, and it may by itself be more or less ineffective. What is needed, after all, is not simply an increase in the funds available for secure investment, but an increase in the money in the hands of industrialists. . . . Firm control, or even nationalization, of the banking system may therefore be required. He then continued:
>
> The prosperity programme should not consist entirely of monetary measures. The Government should make every effort to expand the demand for, as well as the supply of, credit. This should be done by the orthodox method of a public works programme, the encouraging of Government departments and local authorities to push on with construction and development work. . . .

In 1939 Gaitskell went to work for Hugh Dalton at the Ministry of Economic Warfare, never to return to his University College post or to academic economics. During the course of the war he served in a number of increasingly senior positions in the Civil Service but at the war's end he declined a permanent appointment, choosing instead to continue a political career. Gaitskell was elected as the Labour member for South Leeds in 1945 and after a short time in Parliament he was appointed (in 1950) Chancellor of the Exchequer. The Labour Party was defeated in the election of 1951 and Gaitskell became the 'Shadow' Chancellor until December 1955 when, on the retirement of Attlee, he was elected leader of the Parliamentary Labour Party and became Leader of the Opposition. So he remained until his death in January 1963.

Selected Works

1929. *Chartism*. London: Workers' Educational Association.
1933. Four monetary heretics. In *What everyone wants to know about money*, ed. G.D.H. Cole. London: Gollancz.
1935. Financial policy in the transition period. In *New trends in socialism*, ed. G. Catlin. London: Lovat Dickson and Thompson.
1936, 1938. Notes on the period of production. *Zeitschrift für Nationalökonomie* 7(5): (1936) and 9(2): (1938).
1940. *Money and everyday life*. London: Book Service.

References

Williams, P.M. 1982. *Hugh Gaitskell*. Oxford: Oxford University Press.

Galbraith, John Kenneth (1908–2006)

Lester C. Thurow

Keywords

Corporatism; Countervailing power; Galbraith, J. K.; Invisible hand; Planning; Price control; Technical change

JEL Classifications

B31

John Kenneth Galbraith was a paradox. Born in Canada in 1908, he began his professional career armed with a Ph.D. in agricultural economics from the University of California. During the Second World War he was in charge of price controls and immediately after was director of the Strategic Bombing Survey. Later he was in charge of economic affairs in the occupied countries and was awarded the Medal of Freedom for his efforts. He became a Professor of Economics at Harvard, a President of the American Economic Association, and an advisor to presidents and presidential candidates, the latter leading to his appointment as ambassador to India during the Kennedy Administration.

Yet throughout this distinguished career the economics profession moved steadily towards more formal mathematizable models and exhibited less and less interest in old-fashioned political economy, while Galbraith himself never moved an iota in either direction. In the spirit that one might expect from a former editor of *Fortune* magazine, his books were written always in the form of verbally persuasive economic tracts, without a hint of mathematics. His interests were always those of political economy, with political considerations ranking at least as high as, and most often higher than, those of economics.

Perhaps because of his writings on the causes and consequences of the Great Depression in *The Great Crash* (1961) and his successful experience as a price controller during the Second World War, he was never a believer in the wisdom of the invisible hand. If there is an essential theme in his economic writings, it is that the government has a role to play in successful economies.

The Affluent Society (1955) documents the tendency of the invisible hand to promote private splendour and public squalor. Others have made that case (before and since; analytically and verbally), but no one has ever grabbed the public's attention with vivid examples as he did. There were other forces also at work, but much of the effort to improve the quality of the public sector during the 1960s can be traced to his writings. Of course, Galbraith would have wanted the government to go much farther and returned to this theme in *The Culture of Contentment* (1992) and *The Good Society: A Humane Agenda* (1996) which describe a range of interventions to address contemporary problems.

Planning, however, was not just something for the public sector. Planning was essential to the smooth functioning of the private sector. As a result large firms had an important role to play in the private economy. They were not just actual or potential anti-trust threats. In many ways *American Capitalism* (1952b) and its doctrines of countervailing power have come to be the accepted wisdom. Big is no longer automatically bad. Major new government antitrust cases have almost disappeared. That said, *The Economics of Innocent Fraud* (1994a) emphasized Galbraith's concerns late in life about the power of corporate managers to shape society.

In the Galbraith view in *The New Industrial State* (1967a), large firms are essential since they finance much of the research and development that leads to the technical innovations that are necessary to secure a rising standard of living. Technical change had traditionally stood outside of economics as an exogenous force, although with the advent of the new growth economics this is no longer the case. Galbraith placed it where it should be at the centre of his analysis and it led to very different conclusions regarding the role of the large firm. Today it is fashionable to point to the many formerly small firms that have become technological leaders, but Galbraith

would reply that most of these firms can be shown to have sprung from the laboratories of some large firm or university.

The invisible hand systematically leads to too few resources for the public sector, too few resources for research and development, and poor coordination between firms, but it also, in Galbraith's view, leads to too few resources for the poor. In *Economic Development* (1962), *The Nature of Mass Poverty* (1979a) and *The Voice of the Poor* (1983a) he has systematically argued for public actions to redress the imbalances produced by the market in the distribution of income. He was never a believer in the virtues of 'trickle down'. And as the percentage of total income going to the bottom 40 per cent of the population fell in the mid-1980s under the impact of America's current experiment with benign neglect, he could claim vindication for his earlier arguments, as he could have in the last decade of his life.

The result was an economist out of the mainstream of economic thought, but in the mainstream of economic events.

Selected Works

1952a. *A theory of price control.* Cambridge, MA: Harvard University Press.

1952b. *American capitalism.* Boston: Houghton Mifflin.

1955. *The affluent society.* Boston: Houghton Mifflin.

1961. *The great crash.* Boston: Houghton Mifflin.

1962. *Economic development.* Cambridge, MA: Harvard University Press.

1967a. *The new industrial state.* Boston: Houghton Mifflin.

1967b. *How to get out of Viet Nam.* New York: New American Library.

1969a. *Ambassador's journal.* Boston: Houghton Mifflin.

1969b. *How to control the military.* New York: Doubleday.

1973a. *A China passage.* Boston: Houghton Mifflin.

1973b. *Economics and the public purpose.* Boston: Houghton Mifflin.

1975. *Money, whence it came, where it went*. Boston: Houghton Mifflin.

1977. *The age of uncertainty.* Boston: Houghton Mifflin.

1979a. *The nature of mass poverty.* Cambridge, MA: Harvard University Press.

1979b. *Annals of an abiding liberal.* Boston: Houghton Mifflin.

1981. *Life in our times.* Boston: Houghton Mifflin.

1983a. *The voice of the poor.* Cambridge, MA: Harvard University Press.

1983b. *The anatomy of poverty.* Boston: Houghton Mifflin.

1987. *Economics in perspective: A critical history.* Boston: Houghton Mifflin.

1992. *The culture of contentment.* Boston: Houghton Mifflin.

1993. *A short history of financial euphoria. Boston: Houghton Mifflin.*

1994a. *The economics of innocent fraud,* Boston: Houghton Mifflin.

1994b. *A journey through economic time.* Boston: Houghton Mifflin.

1996. *The good society: A humane agenda*. Boston: Houghton Mifflin.

2005. A comprehensive biography by R. Parker, *John Kenneth Galbraith: His life, his politics, his economics,* New York: Farrar, Straus and Giroux.

Gale, David (1921–2008)

Joel Sobel

Abstract

This article reviews the research of David Gale, who made lasting contributions to game theory, general equilibrium theory, and growth theory. In addition to his influence on the development of economic theory, his work has had important implications for many branches of mathematics and on mathematical education.

Keywords

Assignment problem; Competitive equilibrium; Convexity; Debreu, G.; Existence of competitive equilibrium; Gale, David; Gale–Shapley algorithm; Game theory; Global univalence; Kuhn, H.; Linear inequalities; Matching; Mathsite; Nikaido, H.; Ramsey, F.; Shapley, L.; Tucker, A.W.; von Neumann, John; Zero-sum game

JEL Classifications

B31

David Gale was born in New York on 13 December 1921, and died in Berkeley, California on 7 March 2008. He received an undergraduate degree from Swarthmore and a Master's degree from the University of Michigan before earning a Ph.D. in Mathematics at Princeton. It was at Michigan, under the influence of Professor Norman Steenrod, that Gale decided to give up his study of physics and pursue a Ph. D. in mathematics. He taught at Brown University from 1950 through 1965 and then joined the faculty at the University of California, Berkeley. His principal appointment was in the Mathematics Department, but he maintained affiliations with the departments of Economics and Industrial Engineering.

Gale won wide recognition for his research. His awards included a Fulbright research fellowship, two Guggenheim fellowships, elections to the American Academy of Arts and Science and the National Academy of Science, the Lester Ford Prize (for outstanding mathematical exposition), the John von Neumann Theory Prize (for fundamental contributions to operations research), and the Pirelli International Award (for the Internet Mathematics Museum 'MathSite').

Mukul Majumdar (1992) edited the volume *Equilibrium and Dynamics: Essays in Honour of David Gale.* The *International Journal of Game Theory,* volume 36, Numbers 3–4, March 2008 contains a collection of papers dedicated to David Gale on the occasion of his 85th birthday. This volume was edited by Marilda Sotomayor, who had also organized a scientific day in David's honour during the 18th Summer Festival on Game Theory in Stony Brook, 12/13 July 2007. Special issues of *Games and Economic Behavior* and *The Mathematical Intelligencer* are in preparation.

Gale lived in Berkeley, California and Paris, France with his partner Sandra Gilbert, a renowned feminist literary scholar and poet. Her 2000 book of poetry *Kissing the Bread* included a section of poems she wrote for Gale called 'When she was kissed by the mathematician'. He had three daughters and two grandsons. Julie Gale, his former wife and the mother of his daughters, died in February 2008.

Linear Inequalities

As a graduate student at Princeton, David Gale worked with classmate Harold Kuhn on a research project supervised by Professor Albert Tucker. At the time, there was considerable excitement about the new fields of zero-sum game theory and linear programming, but the mathematics of linear inequalities had not been developed. Existing proofs of the minimax theorem of zero-sum game theory required fixed-point arguments and did not make the relationship between the theory and linear inequalities explicit. The project led to important results that identified the deep connections between the two new areas. Gale et al. (1951) contained the first complete proof of the duality theorem of linear programming, and used the theorem to prove the minimax theorem of zero-sum, two-person game theory. This paper uses convex analysis rather than fixed-point arguments to prove the minimax theorem, and implicitly provided a computational foundation for equilibrium points in zero-sum games.

Gale's book *The Theory of Linear Economic Models* (1960) contains central results on the theory of linear inequalities, including Gale, Kuhn and Tucker (1951) and Gale's extension of von Neumann's model of an expanding economy (1956a). It discusses Dantzig's simplex algorithm and gives an economic interpretation to canonical problems. The book also contains a concise and

elementary introduction to the theory of linear inequalities (including a proof of the separating hyperplane theorem for convex polytopes), a chapter containing essential results on non-negative matrices, and a clean treatment of dynamic linear models of growth.

Largely in recognition of their joint work, Gale, Kuhn and Tucker won the 1980 von Neumann prize for work they began in the late 1940s. Their citation stated that they 'played a seminal role in laying the foundations of game theory, linear and nonlinear programming – work that continues to be of fundamental importance to modern operations research and management science'.

Infinite Games

Early work on non-cooperative game theory concentrated on two-player, zero-sum games. When players had finitely many pure strategies, these games were well understood. All two-player finite zero-sum games have a value, and by playing to maximize their minimum payoff, a player could guarantee this value independent of his or her opponent's strategy. Gale and Stewart (1953) studied a class of infinite zero-sum games and demonstrated that the minmax theorem need not hold in this more general setting. The paper examines the simplest possible infinite zero-sum game. In the two-player game of perfect information, the players take turns naming binary digits, which can be thought of as the binary expansion of a number between zero and one. The first player wins if the expansion is an element of a pre-specified set. Otherwise, the second player wins. Gale and Stewart show that basic results from finite games hold if the prespecified set is closed, but that in general the game does not have a value.

The class of infinite games introduced by Gale and Stewart has had broad implications for mathematics. Gale and Stewart's result led to research that identified a general set of games that do have values, culminating in a theorem of Martin (1975). The fact that some games do not have values led to developments in set theory (Mycielski 1964).

Growth

Gale (1956a, 1960, chapter 8, section 5) generalizes and simplifies the von Neumann (1937/1946) model of an expanding economy in what in now called the von Neumann–Gale model of growth. Gale made two essential additions to the original model. He substituted von Neumann's requirement that each production process involves each good in the economy (as either input or output), with a weaker, more plausible condition.

Gale (1967) provides the definitive treatment of the multi-good Ramsey problem, which is a generalization of the linear von Neumann–Gale model. An agent starts with a given endowment, which must be allocated between immediate consumption and investment. What the agent invests is transformed, via a given technology, into the next period's endowment, which again may be allocated between immediate consumption and investment. The process continues indefinitely. The agent cares about the (undiscounted) sum of utility received from consumption. The problem is to find the appropriate definition for optimality and to characterize optimal consumption paths. Gale presents an appropriate optimality condition, provides conditions under which optimal paths exist, and characterizes these paths in terms of 'turnpike' properties. Roughly speaking, we can construct an optimal program with two phases: a bounded initial transition phase in which the state is built up to (approximate) a sustainable optimal steady state, followed by a program that approximates the best steady state consumption.

General Equilibrium

Gale made several important contributions to the foundations of general equilibrium theory. Indeed, he made basic contributions to the three central issues of the theory: existence, uniqueness and stability.

Gale (1955) contains a result known as the Gale–Debreu–Nikaido Lemma (Debreu 1956; Nikaido 1956) which contains the essential mathematical result needed to prove the existence of market equilibrium. Gale and Nikaido

(1965) proves a theorem on the global univalence of differentiable mappings on R^n. When translated into a general equilibrium context, the theorem gives sufficient conditions for equilibrium prices to be unique (see, for example, Arrow and Hahn 1971, chapter 9).

Gale (1963) provides an early robust example of global instability of the tâtonnement process in general equilibrium.

Gale and Mas-Colell (1975) provides an existence theorem in an economy without ordered preferences.

College Admissions and the Stability of Marriage

Gale's paper with Lloyd Shapley on the stable marriage problem (1962) is his most cited, and probably most influential, work. Detailed overviews of the research appear in Knuth (1976/1997), Gale (2001), Roth (2008) and Roth and Sotomayor (1990).

The short, deceptively simple paper is important for several reasons. The motivation for the problem comes from the real world. Gale (2001) describes how the idea for the problem came from thinking about the college application process. The translation of the practical problem into mathematics captures many important considerations but remains extraordinarily simple. The solution to the problem is not obvious, but is easy to understand. The framework lends itself to modifications that lead to insight into more complicated practical problems.

The basic problem is how to create an assignment of items from one group to items from another. The groups can be men and woman (the marriage problem), workers to jobs (labour-market matching), or students to universities (college admissions). For concreteness, consider the marriage problem, in which it is natural to impose the constraint that there are equal numbers of men and women and the desired matching is one to one. Assume everyone has preferences over potential partners (so each man can order the women from the most preferred to least preferred marriage partner, and likewise each woman can order the men). Finding a match is simple. One can order them by age and match the youngest man to the youngest woman, the second youngest man to the second youngest woman, and so on. Gale and Shapley looked for a matching in which there is no unmatched man and woman who prefer each other to their current partners. If this property failed, you would expect the matching to be unstable. Gale and Shapley show that stable matchings exist, and present a simple algorithm that constructs stable matches.

Starting in 1951, 11 years before the publication of the Gale–Shapley paper, the National Intern Matching Program used an essentially equivalent algorithm to match graduating medical students to hospital residency programmes (see Roth and Sotomayor 1990, pp. 169–170; Roth 2008, appendix for a discussion of the independent development of the matching algorithm). Practical problems in a wide variety of areas (from school assignment to kidney exchange) continue to stimulate the development of matching theory.

Assignment Markets and Auctions

Gale's work shows sensitivity to computational issues. Knowing the connection between zero-sum, two-person game theory and the theory of linear programming combined with computational methods (like the simplex method) provides a tractable method for computing equilibria in two-player games. Demonstrating the equivalence between an equilibrium and the solution to a well-behaved optimization problem is the reason that equilibria in linear economies studied by Eisenberg and Gale (1959) can be found efficiently. Gale's work on markets with indivisible goods is another example of a situation in which Gale adds just enough structure to a general model to obtain strong results.

Shapley and Shubik (1971) introduce the assignment model, a market equilibrium model with indivisible goods. Their model has the structure of a matching game with the added feature that agents can exchange a divisible commodity, money. Demange and Gale (1985) show that this

market inherits many properties from the college admissions problem. Demange, Gale, and Sotomayor (1986) apply the framework to the study of multi-unit auctions and show how to define variations of the Vickrey auction to the multi-good setting.

Other Contributions

Gale (1956b) contains a lasting contribution to the study of convex polyhedra, introducing what are now known as 'Gale transforms' and 'Gale diagrams' (see Grünbaum 2003).

Gale (2009) describes the board games invented by Gale and his contemporaries at Princeton. Gale's article provides an introduction to Bridg-It (or, as Martin Gardner called it, the 'Game of Gale') and also John Nash's game of Hex. Gale (1974) invented the game of Chomp, a simple two-player game of perfect information in which it is easy to show that one player has a winning strategy, but the winning strategy is hard to find in general.

Mathematical Explorations

Gale made examples of beautiful mathematical arguments accessible to a broad audience. Between 1991 and 1996 he wrote a column entitled 'Mathematical explorations' for *The Mathematical Intelligencer.* The columns, collected in a book titled *Tracking the Automatic Ant* (Gale 1998), are in the tradition of Martin Gardner's long-running 'Mathematical games' column in *Scientific American.* He also developed MathSite, a pedagogic website that uses interactive exhibits to illustrate important mathematical ideas. MathSite won the 2007 Pirelli International Award for Science Communication in Mathematics.

See Also

- ▶ Auctions (Applications)
- ▶ Convex Programming
- ▶ Existence of General Equilibrium
- ▶ Game Theory
- ▶ Linear Programming
- ▶ Matching
- ▶ Ramsey Model

Selected Works

1951. (With Harold W. Kuhn, and Albert W. Tucker.) Linear programming and the theory of games. In *Activity analysis of production and allocation,* ed. Tjallings C. Koopmans, ch. 19, 287–297. New York: Wiley.

1953. (With Frank M. Stewart.) Infinite games with perfect information. In *Contributions to the theory of games*, ed. Harold W. Kuhn and Albert W. Tucker. Annals of mathematics studies, vol. 28, 245–266. Princeton: Princeton University Press.

1955. The law of supply and demand. *Mathematica Scandinavica* 3.

1956a. A closed linear model of production. In *Linear inequalities and related systems,* ed. Harold W. Kuhn and Albert W. Tucker. Annals of mathematics studies, ch. 18, vol. 38, 285–303. Princeton: Princeton University Press.

1956b. Neighboring vertices on a convex polyhedron. In *Linear inequalities and related systems,* ed. Harold W. Kuhn and Albert W. Tucker. Annals of mathematics studies, ch. 15, vol. 38, 255–263. Princeton: Princeton University Press.

(With Edmund Eisenberg.) Consensus of subjective probabilities: The parimutuel method. *Annals of Mathematical Statistics* 30: 165–168.

The theory of linear economic models. New York: McGraw-Hill.

(With Lloyd S. Shapley.) College admissions and the stability of marriage. *American Mathematical Monthly* 69: 9–15.

A note on global instability of competitive equilibrium. *Naval Research Logistics Quarterly* 10: 81–87.

1965. (With Hukune Nikaido.) The Jacobian matrix and global univalence of mappings. *Mathematische Annalen* 2: 81–93.

1967. On optimal development in a multi-sector economy. *Review of Economic Studies* 34: 1–18.
A curious Nim-type game. *American Mathematical Monthly* 81: 876–879.
(With Andreu Mas-Colell.) An equilibrium existence theorem for a general model without ordered preferences. *Journal of Mathematical Economics* 2: 9–15.
(With Gabrielle Demange.) The strategy structure of two-sided matching markets. *Econometrica* 53: 873–888.
(With Gabrielle Demange and Marilda Sotomayor.) Multi-item auctions. *Journal of Political Economy* 94: 863–872.
1998. *Tracking the automatic ant*. New York: Springer.
2001. The two-sided matching problem. Origin, development and current issues. *International Game Theory Review* 3: 237–252, June–September.
2009. Topological games at Princeton, a mathematical memoir. *Games and Economic Behavior*.

Acknowledgments I thank Harold Kuhn and Bernhard von Stengel for helpful comments. Sobel (2009) is a more detailed overview of David Gale's research contributions.

Bibliography

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco: Holden–Day.
Debreu, G. 1956. Market equilibrium. *Proceedings of the National Academy of Sciences* 42: 876–878.
Grünbaum, B. 2003. *Convex Polytopes*. 2nd ed. New York: Springer.
Knuth, D.E. 1976. *Mariages Stables et leurs Relations avec d'Autres Problèmes Combinatoires: Introduction à l'Analyse Mathématique des Algorithmes. Montréal: Presses de l'Université de Montréal.* (Trans. as *Stable marriage and its relation to other combinatorial problems: An introduction to the mathamatical analysis of algorithms*. CRM proceedings and lecture notes, vol. 10. *American Mathematical Society*, 1997.)
Majumdar, M., eds. 1992. *Equilibrium and dynamics: Essays in honour of David Gale*. London: Macmillan.
Martin, D.A. 1975. Borel determinacy. *Annals of Mathematics*, 2nd series, 102: 363–371.
Mycielski, J. 1964. On the axiom of determinateness. *Fundamenta Mathematicae* 53: 205–224.
Nikaido, H. 1956. On the classical multilateral exchange problem. *Metroeconomica* 8: 135–145.
Roth, A.E. 2008. Deferred acceptance algorithms: History, theory, practice, and open questions. *International Journal of Game Theory* 36: 537–569.
Roth, A.E., and M. Sotomayor. 1990. *Two-sided matching: A study in game – Theoretic modeling and analysis*, Econometric society monograph series. Cambridge: Cambridge University Press.
Shapley, L.S., and M. Shubik. 1971. The assignment game I: The core. *International Journal of Game Theory* 1: 111–130.
Sobel, J. 2009. Regale: Some memorable results. *Games and Economic Behavior* (forthcoming).
von Neumann, J. 1937. Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktzatzes. *Ergebnisse eines mathematischen Kolloquiums* 8, Wien. [Trans. as A model of general economic equilibrium. *Review of Economic Studies* 13: 1–9, 1945–1946.]

Galiani, Ferdinando (1728–1787)

Filippo Cesarano

Keywords

Balance of payments adjustment; Fiduciary monetary systems; Galiani, F.; Hume, D.; Inflationary expectations; Invisible hand; Labour theory of value; Law, J.; Laws of nature; Money supply; Paradox of value; Quantity theory of money; Scarcity; Targets; Time preference theory of interest; Utility; Value, theory of

JEL Classifications

B31

Galiani was born at Chieti, Italy, on 2 December 1728 and died in Naples on 30 October 1787. At the age of seven he was sent to Naples, where he received a classical education under the supervision of his uncle Celestino Galiani, chief almoner to the king. The young Galiani was in close touch with the cultural circles of the time and was soon introduced to the study of economics. In 1744 he translated some of Locke's writings on money.

One year later he took religious orders. His extensive monetary studies culminated in the publication of *Della moneta* (1751), his main work. In 1759 he was appointed secretary of the Neapolitan embassy in Paris where he lived, almost without interruptions, for about ten years. At the end of his stay he wrote the *Dialogues sur le commerce des bléds* (1770). After his return to Naples, Galiani held several high positions in the civil service and published other essays on policy issues and in fields outside economics (Galiani 1974, 1975).

Most of Galiani's theoretical work can be found in his *Della moneta* (1751), which appeared when he was 22. Despite the variety of topics addressed in the book, the basic contributions concern value and monetary theory. Having defined value as a relationship of subjective equivalence between a quantity of one commodity and a quantity of another, Galiani argues that value depends on utility (*utilitá*) and scarcity (*raritá*) (1751, pp. 36–56). Utility is the property of commodities to procure welfare or happiness. Man does not wish to satisfy only primary wants – like eating, drinking, and sleeping – because, once the latter have been satisfied, several others emerge so that full satisfaction is not attainable. Thus, a non-satiation postulate is assumed to hold. Scarcity refers to the quantity of goods available in the market.

Although the interdependence between price and quantity in the determination of market equilibrium is clearly explained by Galiani (1751, pp. 53–4), together with the concept of demand elasticity with respect to wealth, he states that the value of commodities is given by the quantity of labour. Galiani's stress is on value as a relative notion, not related to the intrinsic properties of commodities (1751, p. 119). This theoretical framework allows him to offer a lucid explanation of the so-called paradox of value; according to Schumpeter (1954, p. 300), he 'carried this analysis to its 18th-century peak'.

The main subject of Galiani's 1751 book, however, is money. In order to analyse the properties of a monetary economy, he inquires into the feasibility of dispensing with the use of money altogether, as in religious communities (1751, pp. 87–91). In a large society, goods could be deposited in public warehouses where each producer would be given a receipt (*bullettino*) stating the quantity of commodities deposited so that he would be entitled to withdraw an equivalent amount of commodities. Relative prices would be fixed by the prince. Yet these receipts are nothing but money; money is the means by which everyone's product is represented. Galiani's analysis foreshadows a basic idea shown by recent research (Ostroy 1973), that is, that money is a mechanism to avoid inconsistent claims on commodities on the part of individuals who are motivated by self-interest (1751, p. 90). This analysis notwithstanding, Galiani vigorously rejects a fiduciary monetary system, likely under the influence of events related to John Law's experience in France. These results provide the basis for his theory of the origin of money (1751, pp. 74–81). Media of exchange were not deliberately introduced by man but emerged because some goods had properties that let them be used as means of payment. Galiani's important insight – that the commodities performing monetary functions should be of uniform quality and easily recognizable in order to bring about the reduction of transaction costs and the production of information – can be found in recent work on the subject (Jones 1976, p. 775).

The validity of the quantity theory of money is taken for granted by Galiani. There is, however, a dynamic process through which equilibrium is attained and during this adjustment period changes in money supply affect the economy (1751, pp. 187–9). The same argument was advanced by David Hume in a celebrated passage (1752, pp. 37–8), one year after the publication of *Della moneta*. Although the inefficacy of expected inflation is clearly stated by Galiani (1751, p. 189), an unexpected increase in prices is thought to bring about benefits and costs. Both are discussed at length, but the analysis is rather poor and marred by inconsistencies.

However, Galiani clearly understands that inflation is a concealed way of levying taxes (1751, pp. 198–9, 203–4, 208) and favours the recourse to such a policy in a critical situation when the benefits will more than offset the eventual costs (Cesarano 1976, 1983).

As regards the international aspects of monetary economics, several passages in *Della moneta* show the basic principles of the theory of balance of payments adjustment, pointing out that money flows should not be tampered with by laws or regulations. Galiani views the balance of payments as an essentially monetary phenomenon and payments imbalances as a necessary event which should never be meddled with. Finally, the rate of interest is defined as the relative price of goods dated at different points in time (1751, pp. 290–1), stressing the role of different degrees of risk. In this analysis, an anticipation of the time preference theory of interest may be found.

Galiani places full trust on the laws of nature which regulate economic phenomena. These laws have universal validity and, like physical laws, can never be violated. Hence, the implementation of policy actions is constrained by the existence of natural laws (Cesarano 1976, section 1). The economic process is guided by a 'supreme Hand' (1751, p. 57) which is the religiously biased counterpart (Galiani was an abbot) of Adam Smith's 'invisible hand' a quarter of a century later. This methodological standpoint can also be found in his later book *Dialogues sur le commerce des bléds* (1770), a discussion of the 1764 French law liberalizing corn exports. The theoretical contributions of this work are not as remarkable as those of *Della moneta*. Nevertheless, the *Dialogues* are to be noted for the rather modern treatment of the principles of economic policy. The latter (1770, pp. 319–23) centres upon the fixing of a target and the choice of the means to achieve it. Galiani stresses the need to avoid abrupt changes in policy and to consider the institutional and political setting before following a specific policy. Although natural laws cannot be violated in the long run and so impose a constraint on policy actions, the latter can be effective in the short run.

Galiani's work on economics reveals a large number of contributions putting him far ahead of his time. Concerning the theory of value, Schumpeter stated:

> ... he [Galiani] displayed sure-footed mastery of analytical procedure and, in particular, neatness in his carefully defined conceptual constructions to a degree that would have rendered superfluous all the 19th-century squabbles – and misunderstandings – on the subject of value had the parties to these squabbles first studied his text, *Della moneta*, 1751. (1954, pp. 300–1)

His analysis of the subject of money embodies a rather coherent theoretical structure showing the basic principles upon which classical monetary theory is built.

Selected Works

1751. *Della moneta*. Introduction by G. Caracciolo, ed. A. Merola. Milan: Feltrinelli, 1963. Reprinted 1780 with the addition of a foreword, 35 notes and an epilogue. Other editions include: P. Custodi, ed., in *Scrittori classici italiani di economia politica*. Parte Moderna, vols. 3 and 4, Milan: Destefanis, 1803; F. Nicolini, ed., Bari: Laterza, 1915. English translation by P.R. Toscano, as *On money*. Ann Arbor: University Microfilms International, 1977. A French partial translation has been edited by G.H. Bousquet and J. Crisafulli, *De la monnaie*. Paris: Riviére, 1955. An English translation of the main passages can be found in A.E. Monroe, *Early economic thought*. Cambridge, MA: Harvard University Press, 1924.

1770. *Dialogues sur le commerce des bléds*, ed. F. Nicolini. Naples: Ricciardi, 1959.

1974. *Nuovi saggi inediti di economia*. Introduction by G. Demaria ed A. Agnati. Padua: Cedam.

1975. *Opere*, ed. F. Diaz, and M. Guerci. Naples: Ricciardi.

Bibliography

Accademia Nazionale dei Lincei. 1975. *Ferdinado Galiani. Quaderno N. 211*. Rome: Accademia Nazionale dei Lincei.

Cesarano, F. 1976. Monetary theory in Ferdinando Galiani's *Della moneta*. *History of Political Economy* 8: 380–399.

Cesarano, F. 1983. The rational expectations hypothesis in retrospect. *American Economic Review* 73: 198–203.

Einaudi, L. 1953. Galiani economista. In *Saggi bibliografici e storici intorno alle dottrine economiche*. Rome: Edizioni di Storia e Letteratura.

Hume, D. 1752. Of money. In *Writings on economics*, ed. E. Rotwein. Madison: University of Wisconsin Press, 1955.
Jones, R.A. 1976. The origin and development of media of exchange. *Journal of Political Economy* 84: 757–775.
Ostroy, J.M. 1973. The information and efficiency of monetary exchange. *American Economic Review* 63: 597–610.
Schumpeter, J.A. 1954. *History of economic analysis*. New York: Oxford University Press.
Tagliacozzo, G., ed. 1937. *Economisti napoletani dei sec. XVII e XVIII*. Bologna: Cappelli.

Game Theory

R. J. Aumann

Abstract

Game theory concerns the behaviour of decision makers whose decisions affect each other. Its analysis is from a rational rather than a psychological or sociological viewpoint. It is indeed a sort of umbrella theory for the rational side of social science, where 'social' is interpreted broadly, to include human as well as non-human players (computers, animals, plants). Its methodologies apply in principle to all interactive situations, especially in economics, political science, evolutionary biology, and computer science. There are also important connections with accounting, statistics, the foundations of mathematics, social psychology, law, business, and branches of philosophy such as epistemology and ethics.

Keywords

Asymmetric information; Aumann, R. J.; Axelrod, R.; Axiomatics; Bargaining; Bounded rationality; Brouwer's fixed-point theorem; Coalitional games; Coalitions; Common knowledge assumption; Competitive equilibrium; Consistency; Continuous games; Cooperative game theory; Cooperative games; Cores; Correlated equilibria; Cost allocation; Differential games; Distributed computing; Duality theorems; Dynamic games; Ethics; Evolutionary economics; Expected utility theory; Extensive form games; Fixed threats; Folk theorem; Game theory; Games of incomplete information; General equilibrium; Gillies, D.; Harsanyi, J.; Impossibility theorem; Imputations; Individual rationality; Kakutani's fixed point theorem; Kernel; Kuhn, H.; Lemke-Howson algorithm; Linear inequalities; Linear programming; Luce, R.; Market games; Mathematical economics; Mathematical programming; Milnor, J.; Minimax theorem; Mixed strategy game; Monopoly; Monopsony; Morgenstern, O.; Nash equilibrium; Nash program; Nash, J.; Net worth; Non-cooperative games; Nucleolus; Oligopoly; Perfect information; Prisoner's dilemma; Probability distributions; Raiffa, H.; Ramsey, F.; Randomization; Refinements of Nash equilibrium; Repeated games; Savage, L.; Selten, R.; Shadow pricing; Shapley value; Shapley, L.; Shubik, M.; Small worlds; Maynard Smith, J.; Solution concepts; Stable set theory; Statistical decision theory; Stochastic games; Strategic equilibrium; Strategic games; Strictly competitive games; Strictly determined games; Subgame perfection; Subgames; Super additivity; Testing; Tit for tat; Transferable utility; Tucker, A.; Uncertainty; Utility functions; Value equivalence principle; von Neumann, J.; Voting games; Weighted voting game; Zermelo's theorem; Zero-sum games

JEL Classifications

C7

Introduction

'Interactive decision theory' would perhaps be a more descriptive name for the discipline usually called game theory. This discipline concerns the behaviour of decision makers (*players*) whose decisions affect each other. As in non-interactive (one-person) decision theory, the analysis is from a rational, rather than a psychological or sociological viewpoint. The term 'game theory' steams

from the formal resemblance of interactive decision problems (*games*) to parlour games such as chess, bridge, poker, monopoly, diplomacy or battleship. The term also underscores the rational, 'cold', calculating nature of the analysis.

The major applications of game theory are to economics, political science (on both the national and international levels), tactical and strategic military problems, evolutionary biology, and, most recently, computer science. There are also important connections with accounting, statistics, the foundations of mathematics, social psychology, and branches of philosophy such as epistemology and ethics. Game theory is a sort of umbrella or 'unified field' theory for the rational side of social science, where 'social' is interpreted broadly, to include human as well as non-human players (computers, animals, plants). Unlike other approaches to disciplines like economics or political science, game theory does not use different, ad-hoc constructs to deal with various specific issues, such as perfect competition, monopoly, oligopoly, international trade, taxation, voting, deterrence, and so on.

Rather, it develops methodologies that apply in principle to *all* interactive situations, then sees where these methodologies lead in each specific application. Often it turns out that there are close relations between results obtained from the general game-theoretic methods and from the more ad-hoc approaches. In other cases, the game-theoretic approach leads to new insights, not suggested by other approaches.

We use a historical framework for discussing some of the basic ideas of the theory, as well as a few selected applications. But the viewpoint will be modern; the older ideas will be presented from the perspective of where they have led. Needless to say, we do not even attempt a systematic historical survey.

1910–1930

During these earliest years, game theory was preoccupied with *strictly competitive* games, more commonly known as *two-person zero-sum* games. In these games, there is no point in cooperation or joint action of any kind: if one outcome is preferred to another by one player, then the preference is necessarily reversed for the other. This is the case for most two-person parlour games, such as chess or two-sided poker; but it seems inappropriate for most economic or political applications. Nevertheless, the study of the strictly competitive case has, over the years, turned out remarkably fruitful; many of the concepts and results generated in connection with this case are in fact much more widely applicable, and have become cornerstones of the more general theory. These include the following:

(i) The *extensive*(or *tree*) *form* of a game, consisting of a complete formal description of how the game is played, with a specification of the sequence in which the players move, what they know at the times they must move, how chance occurrences enter the picture, and the *payoff* to each player at the end of play. Introduced by von Neumann (1928), the extensive form was later generalized by Kuhn (1953), and has been enormously influential far beyond zero-sum theory.

(ii) The fundamental concept of *strategy* (or pure strategy) of a player, defined as a complete plan for that player to play the game, as a function of what he observes during the course of play, about the play of others and about chance occurrences affecting the game. Given a strategy for each player, the rules of the game determine a unique outcome of the game and hence a payoff for each player. In the case of two-person zero-sum games, the sum of the two payoffs is zero; this expresses the fact that the preferences of the players over the outcomes are precisely opposed.

(iii) The strategic (or matrix) form of a game. Given strategies $s^1, \ldots, s^n$ for each of the n players, the rules of the game determine a unique outcome, and hence a payoff $H^i(s^1, \ldots, s^n)$ for each player i. The *strategic* form is simply the function that associates to each profile $s := (s^1, \ldots, s^n)$ of strategies, the *pay off profile*

$$H(s) := \left(H^1(s), \ldots\ldots, H^n(s)\right).$$

For two-person games, the strategic form often appears as a matrix: the rows and columns represent pure strategies of Players 1 and 2 respectively, whereas the entries are the corresponding payoff profiles. For zero-sum games, of course, it suffices to give the payoff to Player 1. It has been said that the simple idea of thinking of a game in its matrix form is in itself one of the greatest contributions of game theory. In facing an interactive situation, there is a great temptation to think only in terms of 'what should I do?'. When one writes down the matrix, one is led to a different viewpoint, one that explicitly takes into account that the other players are also facing a decision problem.

(iv) The concept of *mixed* or *randomized* strategy, indicating that rational play is not in general describable by specifying a single pure strategy. Rather, it is often nondeterministic, with specified probabilities associated with each one of a specified set of pure strategies. When randomized strategies are used, payoff must be replaced by expected payoff. Justifying the use of expected payoff in this context is what led to expected utility theory, whose influence extends far beyond game theory (see *1930–1950*, viii).

(v) The concept of 'individual rationality'. The *security level* of Player i is the amount max min $H^i(s)$ that he can guarantee to himself, independent of what the other players do (here the max is over i's strategies, and the min is over $(n-1)$-tuples of strategies of the players other than i). An outcome is called *individually rational* if it yields each player at least his security level. In the game tic-tac-toe, for example, the only individually rational outcome is a draw; and indeed, it does not take a reasonably bright child very long to learn that 'correct' play in tic- tac-toe always leads to a draw.

Individual rationality may be thought of in terms of pure strategies or, as is more usual, in terms of mixed strategies. In the latter case, what is being 'guaranteed' is not an actual payoff, but an expectation; the word 'guarantee' means that this level of payoff can be attained in the mean, regardless of what the other players do. This 'mixed' security level is always at least as high as the 'pure' one. In the case of tic-tac-toe, each player can guarantee a draw even in the stronger sense of pure strategies. Games like this – i.e. having only one individually rational payoff profile in the 'pure' sense – are called *strictly determined*.

Not all games are strictly determined, not even all two-person zero-sum games. One of the simplest imaginable games is the one that game theorists call 'matching pennies', and children call 'choosing up' ('odds and evens'). Each player privately turns a penny either heads up or tails up. If the choices match, 1 gives 2 his penny; otherwise, 2 gives 1 his penny. In the pure sense, neither player can guarantee more than -1, and hence the game is not strictly determined. But in expectation, each player can guarantee 0, simply by turning the coin heads up or tails up with 1/2 – 1/2 probabilities. Thus (0, 0) is the only payoff profile that is individually rational in the mixed sense. Games like this – i.e. having only one individually rational payoff profile in the 'mixed' sense – are called *determined*. In a determined game, the (mixed) security level is called the *value*, strategies guaranteeing it *optimal*.

(vi) *Zermelo's theorem*. The very first theorem of Game Theory (Zermelo 1913) asserts that chess is strictly determined. Interestingly, the proof does not construct 'correct' strategies explicitly; and indeed, it is not known to this day whether the 'correct' outcome of chess is a win for white, a win for black, or a draw. The theorem extends easily to a wide class of parlour games, including checkers, go, and chinese checkers, as well as less well-known games such as hex and gnim (Gale 1979, 1974); the latter two are especially interesting in that one can use Zermelo's theorem to show that Player 1 can force a win, though the proof is non-constructive, and no winning strategy is in fact known. Zermelo's

theorem does not extend to card games such as bridge and poker, nor to the variant of chess known as kriegs spiel, where the players cannot observe their opponents' moves directly. The precise condition for the proof to work is that the game be a two-person zero-sum game of *perfect information*. This means that there are no simultaneous moves, and that everything is open and 'above-board': at any given time, all relevant information known to one player is known to all players.

The domain of Zermelo's theorem – two-person zero-sum games of perfect information – seems at first rather limited; but the theorem has reverberated through the decades, creating one of the main strands of game theoretic thought. To explain some of the developments, we must anticipate the notion of *strategic equilibrium* (Nash 1951; see *1950–1960*, i). To remove the two-person zero-sum restriction, H.W. Kuhn (1953) replaced the notion of 'correct', individually rational play by that of equilibrium. He then proved that *every n-person game of perfect information has an equilibrium in pure strategies*.

In proving this theorem, Kuhn used the notion of a *subgame* of a game; this turned out crucial in later developments of strategic equilibrium theory, particularly in its economic applications. A subgame relates to the whole game like a subgroup to the whole group or a linear subspace to the whole space; while part of the larger game, it is self-contained, can be played in its own right. More precisely, if at any time, all the players know everything that has happened in the game up to that time, then what happens from then on constitutes a subgame.

From Kuhn's proof it follows that every equilibrium (not necessarily pure) of a subgame can be extended to an equilibrium of the whole game. This, in turn, implies that every game has equilibria that remain equilibria when restricted to any subgame. R. Selten (1965) called such equilibria *subgame perfect*. In games of perfect information, the equilibria that the Zermelo-Kuhn proof yields are all subgame perfect.

But not all equilibria are subgame perfect, even in games of perfect information. Subgame perfection implies that when making choices, a player looks forward and assumes that the choices that will subsequently be made, by himself and by others, will be rational; i.e. in equilibrium. Threats which it would be irrational to carry through are ruled out. And it is precisely this kind of forward-looking rationality that is most suited to economic applications.

Interestingly, it turns out that subgame perfection is not enough to capture the idea of forward-looking rationality. More subtle concepts are needed. We return to this subject below, when we discuss the great flowering of strategic equilibrium theory that has taken place since 1975, and that coincides with an increased preoccupation with its economic applications. The point we wished to make here is that these developments have their roots in Zermelo's theorem.

A second circle of ideas to which Zermelo's theorem led has to do with the foundations of mathematics. The starting point is the idea of a game of perfect information with an infinite sequence of stages. Infinitely long games are important models for interactive situations with an indefinite time horizon – i.e. in which the players act as if there will always be a tomorrow.

To fix ideas, let A be any subset of the unit interval (the set of real numbers between 0 and 1). Suppose two players move alternately, each choosing a digit between 1 and 9 at each stage. The resulting infinite sequence of digits is the decimal expansion of a number in the unit interval. Let G_A be the game in which 1 wins if this number is in A, and 2 wins otherwise. Using Set Theory's 'Axiom of Choice', Gale and Stewart (1953) showed that Zermelo's theorem is false in this situation. One can choose A so that G_A is not strictly determined; that is, against each pure strategy of 1, Player 2 has a winning pure strategy, and against each pure strategy of 2, Player 1 has a winning pure strategy. They also showed that if A is open or closed, then G_A is strictly determined.

Both of these results led to significant developments in foundational mathematics. The axiom of choice had long been suspect in the eyes of mathematicians; the extremely anti-intuitive

nature of the Gale-Stewart non-determinateness example was an additional nail in its coffin, and led to an alternative axiom, which asserts that G_A is strictly determined for every set A. This axiom, which contradicts the axiom of choice, has been used to provide an alternative axiomatization for set theory (Mycielski and Steinhaus 1964), and this in turn has spawned a large literature (see Moschovakis 1980, 1983). On the other hand, the positive result of Gale and Stewart was successively generalized to wider and wider families of sets A that are 'constructible' in the appropriate sense (Wolfe 1955; Davis 1964), culminating in the theorem of Martin (1975), according to which G_A is strictly determined whenever A is a Borel set.

Another kind of perfect information game with infinitely many stages is the *differential game*. Here time is continuous but usually of finite duration; a decision must be made at each instant, so to speak. Typical examples are games of pursuit. The theory of differential games was first developed during the 1950s by Rufus Isaacs at the Rand Corporation; his book on the subject was published in 1965, and since then the theory has proliferated greatly. A differential game need not necessarily be of perfect information, but very little is known about those that are not. Some economic examples may be found in Case (1979).

(vii) *The minimax theorem*. The minimax theorem of von Neumann (1928) asserts that every two-person zero-sum game with finitely many pure strategies for each player is determined; that is, when mixed strategies are admitted, it has precisely one individually rational payoff vector. This had previously been verified by E. Borel (e.g. 1924) for several special cases, but Borel was unable to obtain a general proof. The theorem lies a good deal deeper than Zermelo's, both conceptually and technically.

For many years, minimax was considered the elegant centre piece of game theory. Books about game theory concentrated on two-person zero-sum games in strategic form, often paying only desultory attention to the non-zero sum theory. Outside references to game theory often gave the impression that non-zero sum games do not exist, or at least play no role in the theory.

The reaction eventually set in, as it was bound to. Game theory came under heavy fire for its allegedly exclusive concern with a special case that has little interest in the applications. Game theorists responded by belittling the importance of the minimax theorem. During the fall semester of 1964, the writer of these lines gave a beginning course in Game Theory at Yale University, without once even mentioning the minimax theorem.

All this is totally unjustified. Except for the period up to 1928 and a short period in the late 1940s, game theory was never exclusively or even mainly concerned with the strictly competitive case. The forefront of research was always in n-person or non-zero sum games. The false impression given of the discipline was due to the strictly competitive theory being easier to present in books, more 'elegant' and complete. But for more than half a century, that is not where most of the action has been.

Nevertheless, it is a great mistake to belittle minimax. While not the centre piece of game theory, it *is* a vital cornerstone. We have already seen how the most fundamental concepts of the general theory – extensive form, pure strategies, strategic form, randomization, utility theory – were spawned in connection with the minimax theorem. But its importance goes considerably beyond this.

The fundamental concept of non-cooperative n-person game theory – the strategic equilibrium of Nash (1951) – is an outgrowth of minimax, and the proof of its existence is modelled on a previously known proof of the minimax theorem. In cooperative n-person theory, individual rationality is used to define the set of *imputations*, on which much of the cooperative theory is based. In the theory of repeated games, individual rationality also plays a fundamental role.

In many areas of interest – stochastic games, repeated games of incomplete information, continuous games (i.e. with a continuum of pure strategies), differential games, games played by automata, games with vector payoffs – the strictly competitive case already presents a good many of

the conceptual and technical difficulties that are present in general. In these areas, the two-person zero-sum theory has become an indispensable spawning and proving ground, where ideas are developed and tested in a relatively familiar, 'friendly' environment. These theories could certainly not have developed as they did without minimax.

Finally, minimax has had considerable influence on several disciplines outside of game theory proper. Two of these are statistical decision theory and the design of distributed computing systems, where minimax is used for 'worst case' analysis. Another is mathematical programming; the minimax theorem is equivalent to the duality theorem of linear programming, which in turn is closely related to the idea of shadow pricing in economics. This circle of ideas has fed back into game theory proper; in its guise as a theorem about linear inequalities, the minimax theorem is used to establish the condition of Bondareva (1963) and Shapley (1967) for the non- emptiness of the core of an n-person game, and the Hart and Schmeidler (1988) elementary proof for the existence of correlated equilibria.

(viii) *Empirics*. The correspondence between theory and observation was discussed already by von Neumann (1928), who observed that the need to randomize arises endogenously out of the theory. Thus the phenomenon of bluffing in poker may be considered a confirmation of the theory. This kind of connection between theory and observation is typical of game theory and indeed of economic theory in general. The 'observations' are often qualitative rather than quantitative; in practice, we do observe bluffing, though not necessarily in the proportions predicted by theory.

As for experimentation, strictly competitive games constitute one of the few areas in game theory, and indeed in social science, where a fairly sharp, unique 'prediction' is made (though even this prediction is in general probabilistic). It thus invites experimental testing. Early experiments failed miserably to confirm the theory; even in strictly determined games, subjects consistently reached individually irrational outcomes. But experimentation in rational social science is subject to peculiar pitfalls, of which early experimenters appeared unaware, and which indeed mar many modern experiments as well. These have to do with the motivation of the subjects, and with their understanding of the situation. A determined effort to design an experimental test of minimax that would avoid these pitfalls was recently made by B. O'Neill (1987); in these experiments, the predictions of theory were confirmed to within less than 1%.

G

1930–1950

The outstanding event of this period was the publication, in 1944, of the *Theory of Games and Economic Behavior* by John von Neumann and Oskar Morgenstern. Morgenstern was the first economist clearly and explicitly to recognize that economic agents must take the interactive nature of economics into account when making their decisions. He and von Neumann met at Princeton in the late 1930s, and started the collaboration that culminated in the *Theory of Games*. With the publication of this book, game theory came into its own as a scientific discipline.

In addition to expounding the strictly competitive theory described above, the book broke fundamental new ground in several directions. These include the notion of a cooperative game, its coalitional form, and its von Neumann-Morgenstern stable sets. Though axiomatic expected utility theory had been developed earlier by Ramsey (1931), the account of it given in this book is what made it 'catchon'. Perhaps most important, the book made the first extensive applications of game theory, many to economics.

To put these developments into their modern context, we discuss here certain additional ideas that actually did not emerge until later, such as the core, and the general idea of a solution concept. At the end of this section we also describe some developments of this period not directly related to the book, including games with a continuum of strategies, the computation of minimax strategies,

and mathematical advances that were instrumental in later work.

(i) *Cooperative games*. A game is called *cooperative* if commitments – agreements, promises, threats – are fully binding and enforceable (Harsanyi 1966, p. 616). It is called *non-cooperative* if commitments are not enforceable, even if pre-play communication between the players is possible. (For motivation, see *1950–1960*, iv.)

Formally, cooperative games may be considered a special case of non-cooperative games, in the sense that one may build the negotiation and enforcement procedures explicitly into the extensive form of the game. Historically, however, this has not been the mainstream approach. Rather, cooperative theory starts out with a formalization of games (the coalitional form) that abstracts away altogether from procedures and from the question of how each player can best manipulate them for his own benefit; it concentrates, instead, on the possibilities for agreement. The emphasis in the non-cooperative theory is on the individual, on what strategy he should use. In the cooperative theory it is on the group: What coalitions will form? How will they divide the available payoff between their members?

There are several reasons that cooperative games came to be treated separately. One is that when one does build negotiation and enforcement procedures explicitly into the model, then the results of a non-cooperative analysis depend very strongly on the precise form of the procedures, on the order of making offers and counteroffers, and so on. This may be appropriate in voting situations in which precise rules of parliamentary order prevail, where a good strategist can indeed carry the day. But problems of negotiation are usually more amorphous; it is difficult to pin down just what the procedures are. More fundamentally, there is a feeling that procedures are not really all that relevant; that it is the possibilities for coalition forming, promising and threatening that are decisive, rather than whose turn it is to speak.

Another reason is that even when the procedures are specified, non-cooperative analyses of a cooperative game often lead to highly non-unique results, so that they are often quite inconclusive.

Finally, detail distracts attention from essentials. Some things are seen better from a distance; the Roman camps around Metzada are indiscernible when one is in them, but easily visible from the top of the mountain. The coalitional form of a game, by abstracting away from details, yields valuable perspective.

The idea of building non-cooperative models of cooperative games has come to be known as the *Nash program* since it was first proposed by John Nash (1951). In spite of the difficulties just outlined, the programme has had some recent successes (Harsanyi 1982; Harsanyi and Selten 1972; Rubinstein 1982). For the time being, though, these are isolated; there is as yet nothing remotely approaching a general theory of cooperative games based on non-cooperative methodology.

(ii) A *game in coalitional form*, or simply *coalitional game* is a function v associating a real number $v(\mathrm{S})$ with each subset S of a fixed finite set I, and satisfying $v(\varnothing) = 0$ (ϕ denotes the empty set).The members of I are called *players*, the subsets S of I *coalitions* and $v(\mathrm{S})$ is the *worth* of S.

Some notation and terminology: The number of elements in a set S is denoted |S|. A *profile* (of strategies, numbers, etc.) is a function on I (whose values are strategies, numbers, etc.). If x is a profile of numbers and S a coalition, we write $x(\mathrm{S}) := \sum_i \in S^{x^i}$. An example of a coalitional game is the *three-person voting game*; here $|\mathrm{I}| = 3$, and $v(S) = 1$ or 0 according as to whether $|\mathrm{S}| \geq 2$ or not. A coalition S is called *winning* if $v(\mathrm{S}) = 1$, *losing* if $v(\mathrm{S}) = 0$. More generally, if w is a profile of non-negative numbers (*weights*) and q (the *quota*) is positive, define the *weighted voting game* v by $v(\mathrm{S}) = 1$ if $w(\mathrm{S}) \geq q$, and $v(\mathrm{S}) = 0$ otherwise. An example is a parliament with several parties. The players are the parties, rather than the individual members of parliament, w^i is the number of seats held by party i, and q is the number of votes necessary to form a government (usually a simple majority of the

parliament). The weighted voting game with quota q and weights w^i is denoted [q; w]; e.g., the three-person voting game is [2; 1, 1, 1].

Another example of a coalitional game is a *market game*. Suppose there are l natural resources, and a single consumer product, say 'bread', that may be manufactured from these resources. Let each player i have an endowment e^i of resources (an l-vector with non-negative coordinates), and a concave production function u^i that enables him to produce the amount $u^i(x)$ of bread given the vector $x = (x_1,\ldots,x_1)$ of resources. Let v(S) be the maximum amount of bread that the coalition S can produce; it obtains this by redistributing its resources among its members in a manner that is most efficient for production, i.e.

$$v(S) = \max\left\{ \sum_{i \in S} u^i(x^i) : \sum_{i \in S} x^i = \sum_{i \in S} e^i \right\}$$

where the x^i are restricted to have non-negative coordinates.

These examples illustrate different interpretations of coalitional games. In one interpretation, the payoff is in terms of some single desirable physical commodity, such as bread; v(S) represents the maximum total amount of this commodity that the coalition S can procure for its members, and it may be distributed among the members in any desired way. This is illustrated by the above description of the market game.

Underlying this interpretation are two assumptions. First, that of *transferable utility* (TU): that the payoff is in a form that is freely transferable among the players. Second, that of *fixed threats*: that S can obtain a maximum of v(S) no matter what the players outside of S do.

Another interpretation is that v(S) represents some appropriate index of S's strength (if it forms). This requires neither transferable utility nor fixed threats. In voting games, for example, it is natural to define v(S) = 1 if S is a winning coalition (e.g. can form a government or ensure passage of a bill), 0 if not. Of course, in most situations represented by voting games, utility is not transferable.

Another example is a market game in which the x^i are consumption goods rather than resources. Rather than bread, $\sum_{i \in S} u^i(x^i)$ may represent a social welfare function such as is often used in growth or taxation theory. While v (S) cannot then be divided in an arbitrary way among the members of S, it still represents a reasonable index of S's strength. This is a situation with fixed threats but without TU.

Von Neumann and Morgenstern considered strategic games with transferable payoffs, which is a situation with TU but without fixed threats. If the profile s of strategies is played, the coalition S may divide the amount $\sum_{i \in S} H^i(s)$ among its members in any way it pleases. However, what S gets depends on what players outside S do. Von Neumann and Morgenstern defined v(S) as the maxmin payoff of S in the two-person zero-sum game in which the players are S and I = S, and the pay off to S is $\sum_{i \in S} H^i(s)$; i.e., as the expected payoff that S can assure itself (in mixed strategies), no matter what the others do. Again, this is a reasonable index of S's strength, but certainly not the only possible one.

We will use the term *TU coalitional game* when referring to coalitional games with the TU interpretation.

In summary, the coalitional form of a game associates with each coalition S a single number v(S), which in some sense represents the total payoff that that coalition can get or may expect. In some contexts, v(S) fully characterizes the possibilities open to S; in others, it is an index that is indicative of S's strength.

(iii) *Solution concepts*. Given a game, what outcome may be expected? Most of game theory is, in one way or another, directed at this question. In the case of two-person zero-sum games, a clear answer is provided: the unique individually rational outcome. But in almost all other cases, there is no unique answer. There are different criteria, approaches, points of view, and they yield different answers.

A *solution concept* is a function (or correspondence) that associates outcomes, or sets of

outcomes, with games. Usually an 'outcome' may be identified with the profile of payoffs that outcome yields to the players, though sometimes we may wish to think of it as a strategy profile.

Of course a solution concept is not just any such function or correspondence, but one with a specific rationale; for example, the strategic equilibrium and its variants for strategic form games, and the core, the von Neumann-Morgenstern stable sets, the Shapley value and the nucleolus for coalitional games. Each represents a different approach or point of view.

What will 'really' happen? Which solution concept is 'right'? None of them; they are indicators, not predictions. Different solution concepts are like different indicators of an economy; different methods for calculating a price index; different maps (road, topo, political, geologic, etc., not to speak of scale, projection, etc.); different stock indices (Dow Jones, Standard and Poor's NYSE, etc., composite, industrials, utilities, etc.); different batting statistics (batting average, slugging average, RBI, hits, etc.); different kinds of information about rock climbs (arabic and roman difficulty ratings, route maps, verbal descriptions of the climb, etc.); accounts of the same event by different people or different media; different projections of the same three-dimensional object (as in architecture or engineering). They depict or illuminate the situation from different angles; each one stresses certain aspects at the expense of others.

Moreover, solution concepts necessarily leave out altogether some of the most vital information, namely that not entering the formal description of the game. When applied to a voting game, for example, no solution concept can take into account matters of custom, political ideology, or personal relations, since they don't enter the coalitional form. That does not make the solution useless. When planning a rock climb, you certainly want to take into account a whole lot of factors other than the physical characteristics of the rock, such as the season, the weather, your ability and condition, and with whom you are going. But you also do want to know about the ratings.

A good analogy is to distributions (probability, frequency, population, etc.). Like a game, a distribution contains a lot of information; one is overwhelmed by all the numbers. The median and the mean summarize the information in different ways; though other than by simply stating the definitions, it is not easy to say how. The definitions themselves do have a certain fairly clear intuitive content; more important, we gain a feeling for the relation between a distribution and its median and mean from experience, from working with various specific examples and classes of examples over the course of time.

The relationship of solution concepts to games is similar. Like the median and the mean, they in some sense summarize the large amount of information present in the formal description of a game. The definitions themselves have a certain fairly clear intuitive content, though they are not predictions of what will happen. Finally, the relations between a game and its core, value, stable sets, nucleolus, and so on is best revealed by seeing where these solution concepts lead in specific games and classes of games.

(iv) *Domination, the core and imputations*. Continuing to identify 'outcome' with 'payoff profile', we call an outcome y of a game *feasible* if the all-player set I can achieve it. An outcome x *dominates* y if there exists a coalition S that can achieve at least its part of x, and each of whose members prefers x to y; in that case we also say that S can *improve upon* y. The *core* of a game is the set of all feasible outcomes that are not dominated.

In a TU coalitional game v, feasibility of x means $x(I) \leq v(I)$, and x dominating y via S means that $x(S) \leq v(S)$ and $x^i > y^i$ for all i in S. The core of v is the set of all feasible y with $y(S) \geq v(S)$ for all S.

At first, the core sounds quite compelling; why should the players be satisfied with an outcome that some coalition can improve upon? It becomes rather less compelling when one realizes that many perfectly ordinary games have empty cores, i.e. every feasible outcome can be improved

upon. Indeed, this is so even in as simple a game as the three-person voting game.

For a coalition S to improve upon an outcome, players in S must trust each other; they must have faith that their comrades inside S will not desert them to make a coalition with other players outside S. In a TU 3-person voting game, $y := (1/3; 1/3; 1/3)$ is dominated via $\{1, 2\}$ by $x := (1/2, 1/2, 0)$. But 1 and 2 would be wise to view a suggested move from y to x with caution. What guarantee does 1 have that 2 will really stick with him and not accept offers from 3 to improve upon x with, say, (0, 2/3, 1/3)? For this he must depend on 2's good faith, and similarly 2 must depend on1's.

There are two exceptions to this argument, two cases in which domination does not require mutual trust. One is when S consists of a single player. The other is when $S = I$, so that there is no one outside S to lure one's partners away.

The requirement that a feasible outcome y be undominated via one-person coalitions (*individual rationality*) and via the all-person coalition (*efficiency or Pareto optimality*) is thus quite compelling, much more so than that it be in the core. Such outcomes are called *imputations*. For TU coalitional games, individual rationality means that $y^i \geq v(i)$ for all i (we do not distinguish between i and $\{i\}$), and efficiency means that $y(\mathrm{I}) = v(\mathrm{I})$. The outcomes associated with most cooperative solution concepts are imputations; the imputations constitute the stage on which most of cooperative game theory is played out.

The notion of core does not appear explicitly in von Neumann and Morgenstern, but it is implicit in some of the discussions of stable sets there. In specific economic contexts, it is implicit in the work of Edgeworth (1881) and Ransmeier (1942). As a general solution concept in its own right, it was developed by Shapley and Gillies in the early 1950s. Early references include Luce and Raiffa (1957) and Gillies (1959).

(v) *Stable sets*. The discomfort with the definition of core expressed above may be stated more sharply as follows. Suppose we think of an outcome in the core as 'stable'. Then we should not exclude an outcome y just because it is dominated by *some* other outcome x; we should demand that x itself be stable. If x is not itself stable, then the argument for excluding y is rather weak; proponents of y can argue with justice that replacing it with x would not lead to a more stable situation, so we may as well stay where we are. If the core were the set of all outcomes not dominated by any element of the core, there would be no difficulty; but this is not so.

Von Neumann and Morgenstern were thus led to the following definition: A set K of imputations is called *stable* if it is the set of all imputations not dominated by any element of K.

This definition guarantees neither existence nor uniqueness. On the face of it, a game may have many stable sets, or it may have none. Most games do, in fact, have many stable sets; but the problem of existence was open for many years. It was solved by Lucas (1969), who constructed a ten-person, TU coalitional game without any stable set. Later, Lucas and Rabie (1982) constructed a 14-person TU coalitional game without any stable set and with an empty core to boot.

Much of the *Theory of Games* is devoted to exploring the stable sets of various classes of TU coalitional games, such as three- and four-person games, voting games, market games, compositions of games, and so on. (If v and w have disjoint player sets I and J, their *composition* u is given by $u(S) := v(\mathrm{S} \cap \mathrm{I}) + w(\mathrm{S} \cap \mathrm{J})$.). During the 1950s many researchers carried forward with great vigour the work of investigating various classes of games and describing their stable sets. Since then work on stable sets has continued unabated, though it is no longer as much in the forefront of game-theoretic research as it was then. All in all, more than 200 articles have been published on stable sets, some 80% of them since 1960. Much of the recent activity in this area has taken place in the Soviet Union.

It is impossible here even to begin to review this large and varied literature. But we do note one characteristic qualitative feature. By definition, a stable set is simply a set of imputations; there is nothing explicit in it about social structure. Yet the

mathematical description of a given stable set can often best be understood in terms of an implicit social structure or form of organization of the players. Cartels, systematic discrimination, groups within groups, all kinds of subtle organizational forms spring to one's attention. These forms are endogenous, they are not imposed by definition, they emerge from the analysis. It is a mystery that just the stable set concept, and it only, is so closely allied with endogenous notions of social structure.

We adduce just one, comparatively simple example. The TU three-person voting game has a stable set consisting of the three imputations (1/2, 1/2, 0), (1/2, 0, 1/2), (0, 1/2, 1/2). The social structure implicit in this is that all three players will *not* compromise by dividing the payoff equally. Rather, one of the three 2-person coalitions will form and divide the payoff equally, with the remaining player being left 'in the cold'. Because any of these three coalitions can form, competition drives them to divide the payoff equally, so that no player will prefer any one coalition to any other.

Another stable set is the interval $\{(\alpha, 1 - \alpha, 0)\}$, where α ranges from 0 to 1.Here Player 3 is permanently excluded from all negotiations; he is 'discriminated against'. Players 1 and 2 divide the payoff in some arbitrary way, not necessarily equally; this is because a coalition with 3 is out of the question, and so competition no longer constrains 1 and 2 in bargaining with each other.

(vi) *Transferable utility.* Though it no longer enjoys the centrality that it did up to about 1960, the assumption of transferable utility has played and continues to play a major role in the development of cooperative game theory. Some economists have questioned the appropriateness of the TU assumption, especially in connection with market models; it has been castigated as excessively strong and unrealistic.

This situation is somewhat analogous to that of strictly competitive games, which as we pointed out above (*1930–1950*, vii), constitute a proving ground for developing and testing ideas that apply also to more general, non-strictly competitive games. The theory of NTU (non-transferable utility) coalitional games is now highly developed (see *1960–1970*, i), but it is an order of magnitude more complex than that of TU games. The TU theory is an excellent laboratory or model for working out ideas that are later applied to the more general NTU case.

Moreover, TU games are both conceptually and technically much closer to NTU games than strictly competitive games are to non-strictly competitive games. A very large part of the important issues arising in connection with non-strictly competitive games do not have any counterpart at all in strictly competitive games, and so simply cannot be addressed in that context. But by far the largest part of the issues and questions arising in the NTU theory do have counterparts in the TU theory; they can at least be addressed and dealt with there.

Almost every major advance in the NTU theory – and many a minor advance as well – has had its way paved by a corresponding advance in the TU theory. Stable sets, core, value, and bargaining set were all defined first for TU games, then for NTU. The enormous literature on the core of a market and the equivalence between it and competitive equilibrium (c.e.) in large markets was started by Martin Shubik (1959a) in an article on TU markets. The relation between the value and c.e. in large markets was also explored first for the TU case (Shapley 1964; Shapley and Shubik 1969b; Aumann and Shapley 1974; Hart 1977a), then for NTU (Champsaur 1975, but written and circulated circa 1970; Aumann 1975; Mas-Colell 1977; Hart 1977b). The same holds for the bargaining set; first TU (Shapley 1984), then NTU (Mas-Colell 1988). The connection between balanced collections of coalitions and the non-emptiness of the core (*1960–1970*, viii) was studied first for TU (Bondareva 1963; Shapley 1967), then for NTU (Scarf 1967; Billera 1970b; Shapley 1973a); this development led to the whole subject of Scarf's algorithm for finding points in the core, which he and others later extended to algorithms for finding market equilibria and fixed points of mappings in general. Games arising from markets were first

abstractly characterized in the TU case (Shapley and Shubik 1969a), then in the NTU case (Billera and Bixby 1973; Mas-Colell 1975). Games with a continuum of players were conceived first in a TU application (Milnor and Shapley 1979, but written and circulated in 1960), then NTU (Aumann 1964). Strategic models of bargaining where time is of the essence were first treated for TU (Rubinstein 1982), then NTU (Binmore 1982). One could go on and on.

In each of these cases, the TU development led organically to the NTU development; it isn't just that the one came before the other. TU is to cooperative game theory what *Drosophila* is to genetics. Even if it had no direct economic interest at all, the study of TU coalitional games would be justified solely by their role as an outstandingly suggestive research tool.

(vii) *Single play.* Von Neumann and Morgenstern emphasize that their analysis refers to 'one-shot' games, games that are played just once, after which the players disperse, never to interact again. When this is not the case, one must view the whole situation – including expected future interactions of the same players – as a single larger game, and it, too, is to be played just once.

To some extent this doctrine appears unreasonable. If one were to take it literally, there would be only one game to analyse, namely the one whose players include all persons ever born and to be born. Every human being is linked to every other through some chain of interactions; no person or group is isolated from any other.

Savage (1954) has discussed this in the context of one-person decisions. In principle, he writes, one should 'envisage every conceivable policy for the government of his whole life in its most minute details, and decide here and now on one policy. This is utterly ridiculous ... ' (p. 16). He goes on to discuss the *small worlds* doctrine, 'the practical necessity of confining attention to, or isolating, relatively simple situations .. .' (p. 82).

To a large extent, this doctrine applies to interactive decisions too. But one must be careful, because here 'large worlds' have qualitative features totally absent from 'small worlds'. We return to this below (*1950–1960*, ii, iii).

(viii) *Expected utility.* When randomized strategies are used in a strategic game, payoff must be replaced by expected payoff (*1910–1930*, iv). Since the game is played only once, the law of large numbers does not apply, so it is not clear why a player would be interested specifically in the mathematical expectation of his pay off.

There is no problem when for each player there are just two possible outcomes, which we may call 'winning' and 'losing', and denominate 1 and 0 respectively. (This involves no zero-sum assumption; e.g. all players could win simultaneously.) In that case the expected payoff is simply the probability of winning. Of course each player wants to maximize this probability, so in that case use of the expectation is justified.

Suppose now that the values of i's payoff function Hi are numbers between 0 and 1, representing win probabilities. Thus, for the 'final' outcome there are still only two possibilities; each pure strategy profile s induces a random process that generates a win for i with probability $H^i(s)$. Then the payoff expectation when randomized strategies are used still represents i's overall win probability.

Now in any game, each player has a most preferred and a least preferred outcome, which we take as a win and a loss. For each payoff h, there is some probability p such that i would as soon get h with certainty as winning with probability p and losing withprobability$1 - p$. If we replace all the h's by the corresponding p's in the payoff matrix, then we are in the case of the previous paragraph, so use of the expected payoff is justified.

The probability p is a function of h, denoted $u^i(h)$, and called i's von Neumann-Morgenstern *utility.* Thus, to justify the use of expectations, each player's pay off must be replaced by its utility.

The key property of the function u^i is that if h and g are random payoffs, then i prefers h to g iff $Eu^i(h) > Eu^i(g)$, where E denotes expectation.

This property continues to hold when we replace u^i by a linear transform of the form $\alpha u^i + \beta$, where α and β are constants with $\alpha > 0$. All these transforms are also called utility functions for i, and any one of them may be used rather than u^i in the pay off matrix.

Recall that a strictly competitive game is defined as a two-person game in which if one outcome is preferred to another by one player, the preference is reversed for the other. Since randomized strategies are admitted, this condition applies also to 'mixed outcomes' (probability mixtures of pure outcomes). From this it may be seen that a two-person game is strictly competitive if and only if, for an appropriate choice of utility functions, the utility payoffs of the players sum to zero in each square of the matrix.

The case of TU coalitional games deserves particular attention. There is no problem if we assume fixed threats and continue to denominate the payoff in bread (see ii). But without fixed threats, the total amount of bread obtainable by a coalition S is a random variable depending on what players outside S do; since this is not denominated in utility, there is no justification for replacing it by its expectation. But if we do denominate payoffs in utility terms, then they cannot be directly transferred. The only way out of this quandary is to assume that the utility of bread is linear in the amount of bread (Aumann 1960). We stress again that no such assumption is required in the fixed threat case.

(ix) *Applications*. The very name of the book, *Theory of Games and Economic Behavior*, indicates its underlying preoccupation with the applications. Von Neumann had already mentioned *Homo Economicus* in his 1928 paper, but there were no specific economic applications there.

The method of von Neumann and Morgenstern has become the archetype of later applications of game theory. One takes an economic problem, formulates it as a game, finds the game-theoretic solution, then translates the solution back into economic terms. This is to be distinguished from the more usual methodology of economics and other social sciences, where the building of a formal model and a solution concept, and the application of the solution concept to the model, are all rolled into one.

Among the applications extensively treated in the book is voting. A qualitative feature that emerges is that many different weight-quota configurations have the same coalitional form; [5; 2, 3, 4] is the same as [2; 1, 1, 1]. Though obvious to the sophisticated observer when pointed out, this is not widely recognized; most people think that the player with weight 4 is considerably stronger than the others (Vinacke and Arkoff 1957). The Board of Supervisors of Nassau County operates by weighted voting; in 1964 there were six members, with weights of 31, 31, 28, 21, 2, 2, and a simple majority quota of 58 (Lucas 1983, p. 188). Nobody realized that three members were totally without influence, that [58; 31, 31, 28, 21, 2, 2]= [2; 1, 1, 1, 0, 0, 0].

In a voting game, a winning coalition with no proper winning subsets is called *minimal winning* (mw). The game $[q; w]$ is *homogeneous* if $w(S) = q$ for all minimal winning S; thus [3; 2, 1, 1, 1] is homogeneous, but [5; 2, 2, 2, 1, 1, 1] is not. A *decisive* voting game is one in which a coalition wins if and only if its complement loses; both the above games are decisive, but [3; 1, 1, 1, 1] is not. TU decisive homogeneous voting games have a stable set in which some mw coalition forms and divides the payoff in proportion to the weights of its members, leaving nothing for those outside. This is reminiscent of some parliamentary democracies, where parties in a coalition government get cabinet seats roughly in proportion to the seats they hold in parliament. But this fails to take into account that the actual number of seats held by a party may well be quite disproportional to its weight in a homogeneous representation of the game (when there is such are presentation).

The book also considers issues of monopoly (or monopsony) and oligopoly. We have already pointed out that stable set theory concerns the endogenous emergence of social structure. In a market with one buyer (monopsonist) and two sellers (duopolists) where supply exceeds demand, the theory predicts that the duopolists

will form a cartel to bargain with the monopsonist. The core, on the other hand, predicts cut-throat competition; the duopolists end up by selling their goods for nothing, with the entire consumer surplus going to the buyer.

This is a good place to point out a fundamental difference between the game- theoretic and other approaches to social science. The more conventional approaches take institutions as given, and ask where they lead. The game theoretic approach asks how the institutions came about, what led to them? Thus general equilibrium theory takes the idea of market prices for granted; it concerns itself with their existence and properties, calculating them, and so on. Game Theory asks, *why* are there market prices? How did they come about? Under what conditions will all traders trade at given prices?

Conventional economic theory has several approaches to oligopoly, including competition and cartelization. Starting with any particular one of these, it calculates what is implied in specific applications. Game theory proceeds differently. It starts with the physical description of the situation only, making no institutional or doctrinal assumptions, then applies a solution concept and sees where it leads.

In a sense, of course, the doctrine is built into the solution concept; as we have seen, the core implies competition, the stable set cartelization. It is not that game theory makes no assumptions, but that the assumptions are of a more general, fundamental nature. The difference is like that between deriving the motion of the planets from Kepler's laws or from Newton's laws. Like Kepler's laws, which apply to the planets only, oligopoly theory applies to oligopolistic markets only. Newton's laws apply to the planets and also to apples falling from trees; stable sets apply to markets and also to voting.

To be sure, conventional economics is also concerned with the genesis of institutions, but on an informal, verbal, ad-hoc level. In game theory, institutions like prices or cartels are outcomes of the formal analysis.

(x) Games with a *continuum of pure strategies* were first considered by Ville (1938), who proved the minimax theorem for them, using an appropriate continuity condition. To guarantee the minimax (security) level, one may need to use a continuum of pure strategies, each with probability zero. An example due to Kuhn (1952) shows that in general one cannot guarantee anything even close to minimax using strategies with finite support. Ville's theorem was extended in the 1950s to strategic equilibrium in non-strictly competitive games.

(xi) *Computing* security levels, and strategies that will guarantee them, is highly non-trivial. The problem is equivalent to that of linear programming, and thus succumbed to the simplex method of George Dantzig (1951a, b).

(xii) The major advance in relevant mathematical methods during this period was *Kakutani's fixed point theorem* (1941). An abstract expression of the existence of equilibrium, it is the vital active ingredient of countless proofs in economics and game theory. Also instrumental in later work were Lyapounov's theorem on the range of a vector measure (1940) and von Neumann's selection theorem (1949).

1950–1960

The 1950s were a period of excitement in game theory. The discipline had broken out of its cocoon, and was testing its wings. Giants walked the earth. At Princeton, John Nash laid the groundwork for the general non-cooperative theory, and for cooperative bargaining theory; Lloyd Shapley defined the value for coalitional games, initiated the theory of stochastic games, co-invented the core with D.B. Gillies, and, together with John Milnor, developed the first game models with continua of players; Harold Kuhn worked on behaviour strategies and perfect recall; Al Tucker discovered the prisoner's dilemma; the Office of Naval Research was unstinting in its support. Three Game Theory conferences were held at Princeton, with the active participation of von Neumann and Morgenstern

themselves. Princeton University Press published the four classic volumes of *Contributions to the Theory of Games*. The Rand Corporation, for many years to be a major centre of game theoretic research, had just opened its doors in Santa Monica. R. Luce and H. Raiffa (1957) published their enormously influential *Games and Decisions*. Near the end of the decade came the first studies of repeated games.

The major applications at the beginning of the decade were to tactical military problems: defense from missiles, Colonel Blotto, fighter-fighter duels, etc. Later the emphasis shifted to deterrence and cold war strategy, with contributions by political scientists like Kahn, Kissinger, and Schelling. In 1954, Shapley and Shubik published their seminal paper on the value of a voting game as an index of power. And in 1959 came Shubik's spectacular rediscovery of the core of a market in the writings of F.Y. Edgeworth (1881). From that time on, economics has remained by far the largest area of application of game theory.

(i) An *equilibrium* (Nash 1951) of a strategic game is a (pure or mixed) strategy profile in which each player's strategy maximizes his payoff given that the others are using their strategies. See the entry on Nash equilibrium, refinements of.

Strategic equilibrium is without doubt the single game theoretic solution concept that is most frequently applied in economics. Economic applications include oligopoly, entry and exit, market equilibrium, search, location, bargaining, product quality, auctions, insurance, principal-agent, higher education, discrimination, public goods, what have you. On the political front, applications include voting, arms control, and inspection, as well as most international political models (deterrence, etc.) Biological applications of game theory all deal with forms of strategic equilibrium; they suggest a simple interpretation of equilibrium quite different from the usual overt rationalism (see *1970–1986*, i). We cannot even begin to survey all this literature here.

(ii) *Stochastic and other dynamic games*. Games played in stages, with some kind of stationary time structure, are called *dynamic*. They include stochastic games, repeated games with or without complete information, games of survival (Milnor and Shapley 1957; Luce and Raiffa 1957; Shubik 1959a, b) or ruin (Rosenthal and Rubinstein 1984), recursive games (Everett 1957), games with varying opponents (Rosenthal 1979), and similar models.

This kind of model addresses the concerns we expressed above (*1930–1950*, vii) about the single play assumption. The present can only be understood in the context of the past and the future: 'Know whence you came and where you are going' (Ethics of the Fathers III:1). Physically, current actions affect not only current payoff but also opportunities and payoffs in the future. Psychologically, too, we learn: past experience affects our current expectations of what others will do, and therefore our own actions. We also teach: our current actions affect others' future expectations, and therefore their future reactions.

Two dynamic models – stochastic and repeated games – have been especially 'successful'. *Stochastic* games address the physical point, that current actions affect future opportunities. A strategic game is played at each stage; the profile of strategies determines both the payoff at that stage and the game to be played at the next stage (or a probability distribution over such games). In the strictly competitive case, with future payoff discounted at a fixed rate, Shapley (1953a) showed that stochastic games are determined; also, that they have optimal strategies that are stationary, in the sense that they depend only on the game being played (not on the history or even on the date). Bewley and Kohlberg (1976) showed that as the discount rate tends to 0 the value tends to a limit; this limit is the same as the limit, as $k \to \infty$, of the values of the k-stage games, in each of which the payoff is the mean payoff for the k stages. Mertens and Neyman (1981) showed that the value exists also in the undiscounted infinite stage game, when payoff is defined by the Cesaro limit (limit, as $k \to \infty$, of

the average payoff in the first k stages). For an understanding of some of the intuitive issues in this work, see Blackwell and Ferguson (1968), which was extremely influential in the modern development of stochastic games.

The methods of Shapley, and of Bewley and Kohlberg, can be used to show that non-strictly competitive stochastic games with fixed discounts have equilibria in stationary strategies, and that when the discount tends to 0, these equilibria converge to a limit (Mertens 1982). But unlike in the strictly competitive case, the payoff to this limit need not correspond to an equilibrium of the undiscounted game (Sorin 1986b). It is not known whether undiscounted non-strictly competitive stochastic games need at all have strategic equilibria.

(iii) *Repeated* games model the psychological, informational side of ongoing relationships. Phenomena like cooperation, altruism, trust, punishment, and revenge are predicted by the theory. These may be called 'subjective informational' phenomena, since what is at issue is information about the behaviour of the players. Repeated games of incomplete information (*1960–1970*, ii) also predict 'objective informational' phenomena such as secrecy, and signalling of substantive information. Both kinds of informational issue are quite different from the 'physical' issues addressed by stochastic games.

Given a strategic game G, consider the game G^∞ each play of which consists of an infinite sequence of repetitions of G. At each stage, all players know the actions taken by all players at all previous stages. The payoff in G^∞ is some kind of average of the stage payoffs; we will not worry about exact definitions here.

The reader is referred to the entry on repeated games. Here we state only one basic result, known as the *Folk Theorem*. Call an outcome (payoff profile) x *feasible* in G if it is achievable by the all-player set when using a correlated randomizing device; i.e. is in the convex hull of the 'pure' outcomes. Call it *strongly individually rational* if no player i can be prevented from achieving x^i by the other players, when they are randomizing independently; i.e. if $x^i \geq \min \max H^i(s)$, where the max is over i's strategies, and the min is over$(n-1)$-tuples of mixed strategies of the others. The Folk Theorem then says that the equilibrium outcomes in the repetition G^∞ *coincide with the feasible and strongly individually rational outcomes in the one-shot game* G.

The authorship of the Folk Theorem, which surfaced in the late 1950s, is obscure. Intuitively, the feasible and strongly individually rational outcomes are the outcomes that could arise in cooperative play. Thus the Folk Theorem points to a strong relationship between repeated and cooperative games. Repetition is a kind of enforcement mechanism; agreements are enforced by 'punishing' deviators in subsequent stages.

(iv) The *Prisoner's Dilemma* is a two-person non-zero sum strategic game with payoff matrix as depicted in Fig. 1. Attributed to A.W. Tucker, it has deservedly attracted enormous attention; it is said that in the social psychology literature alone, over a thousand papers have been devoted to it.

One may think of the game as follows: Each player decides whether he will receive \$1000 or the other will receive \$3000. The decisions are simultaneous and independent, though the players may consult with each other before deciding.

The point is that ordinary rationality leads each player to choose the \$1000 for himself, since he is thereby better off *no matter what the other player does*. But the two players thereby get only \$1000 each, whereas they could have gotten \$3000 each if both had been 'friendly' rather than 'greedy'.

	f	g
f	3,3	0,4
g	4,0	1,1

Game Theory, Fig. 1

The universal fascination with this game is due to its representing, in very stark and transparent form, the bitter fact that when individuals act for their own benefit, the result may well be disaster for all. This principle has dozens of applications, great and small, in everyday life. *People who fail to cooperate for their own mutual benefit are not necessarily foolish or irrational*; they may be acting perfectly rationally. The sooner we accept this, the sooner we can take steps to design the terms of social intercourse so as to encourage cooperation.

One such step, of very wide applicability, is to make available a mechanism for the enforcement of voluntary agreements. 'Pray for the welfare of government, without whose authority, man would swallow man alive' (Ethics of the Fathers III: 2). The availability of the mechanism is itself sufficient; once it is there, the players are naturally motivated to use it. If they can make an *enforceable* agreement yielding (3, 3), they would indeed be foolish to end up with (1, 1). It is this that motivates the definition of a cooperative game (*1930–1950*, i).

The above discussion implies that (g, g) is the unique strategic equilibrium of the prisoner's dilemma. It may also be shown that in any finite repetition of the game, all strategic equilibria lead to a constant stream of 'greedy' choices by each player; but this is a subtler matter than the simple domination argument used for the one-shot case. In the infinite repetition, the Folk Theorem (iii) shows that (3, 3) is an equilibrium outcome; and indeed, there are equilibria that lead to a constant stream of 'friendly' choices by each player. The same holds if we discount future payoff in the repeated game, as long as the discount rate is not too large (Sorin 1986a).

R. Axelrod (1984) has carried out an experimental study of the repeated prisoner's dilemma. Experts were asked to write computer programmes for playing the game, which were matched against each other in a 'tournament'. At each stage, the game ended with a fixed (small) probability; this is like discounting. The most successful program in the tournament turned out to be a 'cooperative' one: Matched against itself, it yields a constant stream of 'friendly' choices; matched against others, it 'punishes' greedy choices. The results of this experiment thus fit in well with received theoretical doctrine.

The design of this experiment is noteworthy because it avoids the pitfalls so often found in game experiments: lack of sufficient motivation and understanding. The experts chosen by Axelrod understood the game as well as anybody. Motivation was provided by the investment of their time, which was much more considerable than that of the average subject, and by the glory of a possible win over distinguished colleagues. Using computer programmes for strategies presaged important later developments (*1970–1986*, iv).

Much that is fallacious has been written on the one-shot prisoner's dilemma. It has been said that for the reasoning to work, pre-play communication between the players must be forbidden. This is incorrect. The players can communicate until they are blue in the face, and agree solemnly on (f,f); when faced with the actual decision, rational players will still choose g. It has been said that the argument depends on the notion of strategic equilibrium, which is open to discussion. This too is incorrect; the argument depends only on strong domination, i.e. on the simple proposition that people always prefer to get another $1000. 'Resolutions' of the 'paradox' have been put forward, suggesting that rational players will play f after all; that my choosing f has some kind of 'mirror' effect that makes you choose it also. Worse than just nonsense, this is actually vicious, since it suggests that the prisoner's dilemma does not represent a real social problem that must be dealt with.

Finally, it has been said that the experimental evidence – Axelrod's and that of others – contradicts theory. This too is incorrect, since most of the experimental evidence relates to repeated games, where the friendly outcome is perfectly consonant with theory; and what evidence there is in one-shot games does point to a preponderance of 'greedy' choices. It is true that in long finite repetitions, where the only equilibria are greedy, most experiments nevertheless point to the friendly outcome; but fixed finite repetitions are somewhat artificial, and besides, this finding,

too, can be explained by theory (Neyman 1985a, b; see *1970–1986*, iv).

(v) We turn now to cooperative issues. A model of fundamental importance is the *bargaining problem* of Nash (1950). Formally, it is defined as a convex set C in the Euclidean plane, containing the origin in its interior. Intuitively, two players bargain; they may reach any agreement whose payoff profile is in C; if they disagree, they get nothing. Nash listed four *axioms* – conditions that a reasonable compromise solution might be expected to satisfy – such as symmetry and efficiency. He then showed that there is one and only one solution satisfying them, namely the point x in the non-negative part of C that maximizes the product x^1x^2. An appealing economic interpretation of this solution was given by Harsanyi (1956).

By varying the axioms, other authors have obtained different solutions to the bargaining problem, notably Kalai and Smorodinski (1975) and Maschler and Perles (1981). Like Nash's solution, each of these is characterized by a formula with an intuitively appealing interpretation.

Following work of A. Rubinstein (1982) and K. Binmore (1982) constructed an explicit bargaining model which, when analyzed as a non-cooperative strategic game, leads to Nash's solution of the bargaining problem. This is an instance of a successful application of the 'Nash program' (see *1930–1950*, vi). Similar constructions have been made for other solutions of the bargaining problem.

An interesting qualitative feature of the Nash solution is that it is very sensitive to risk aversion. A risk loving or risk neutral bargainer will get a better deal than a risk averse one; this is so even when there are no overt elements of risk in the situation, nothing random. The very willingness to take risks confers an advantage, though in the end no risks are actually taken.

Suppose, for example, that two people may divide \$600 in any way they wish; if they fail to agree, neither gets anything. Let their utility functions be $u^1(\$x) = x$ and $u^2(\$x) = \sqrt{x}$, so that 1 is risk neutral, 2 risk averse. Denominating the payoffs in utilities rather than dollars, we find that the Nash solution corresponds to a dollar split of \$400–\$200 in favour of the risk neutral bargainer.

This corresponds well with our intuitions. A fearful, risk averse person will not bargain well. Though there are no overt elements of risk, no random elements in the problem description, the bargaining itself constitutes a risk. A risk averse person is willing to pay, in terms of a less favourable settlement, to avoid the risk of the other side's being adamant, walking away, and soon.

(vi) The *value* (Shapley 1953b) is a solution concept that associates with each coalitional game v a unique outcome φv. Fully characterized by a set of axioms, it may be thought of as a reasonable compromise or arbitrated outcome, given the power of the players. Best, perhaps, is to think of it simply as an index of power, or what comes to the same thing, of social productivity (see Shapley value).

It may be shown that Player i's value is given by

$$\varphi^i v = (1/n!) \sum v^i(S^i_R),$$

where Σ ranges over all $n!$ orders on the set I of all players, $S^i|_R$ is the set of players up to and including i in the order R, and v^i(S) is the *contribution* $v(S) - v(S = i)$ of i to the coalition S; note that this implies linearity of φv in v. In words, $\varphi^i v$ is i's mean contribution when the players are ordered at random; this suggests the social productivity interpretation, an interpretation that is reinforced by the following remarkable theorem (Young 1985): Let ψ be a mapping from games v to efficient outcomes ψv, that is symmetric among the players in the appropriate sense. Suppose $\psi^i v$ depends only on the 2^{n-1} contributions v^i(S), and monotonically so. Then ψ must be the value φ. In brief, if it depends on the contributions only, it's got to be the value, even though we don't assume linearity to start with.

An intuitive feel for the value may be gained from examples. The value of the three-person

voting game is (1/3, 1/3, 1/3), as is suggested by symmetry. This is not in the core, because {1, 2} can improve upon it. But so can {1, 3} and {2, 3}; starting from (1/3, 1/3, 1/3), the players might be well advised to leave things as they are (see 1930–1950, iv). Differently viewed, the symmetric stable set predicts one of the three outcomes (1/2, 1/2, 0), (1/2, 0, 1/2), (0, 1/2, 1/2). Before the beginning of bargaining, each player may figure that his chances of getting into a ruling coalition are 2/3, and conditional on this, his payoff is 1/2. Thus his 'expected outcome' is the value, though in itself, this outcome has no stability.

In the homogenous weighted voting game [3; 2, 1, 1, 1], the value is (1/2, 1/6, 1/6, 1/6); the large player gets a disproportionate share, which accords with intuition: '1'union fait la force.'

Turning to games of economic interest, we model the market with two sellers and one buyer discussed above (*1930–1950*, ix) by the TU weighted voting game [3; 2, 1, 1]. The core consists of the unique point (1, 0, 0), which means that the sellers must give their merchandise, for nothing, to the buyer. While this has clear economic meaning – cutthroat competition – it does not seem very reasonable as a compromise or an index of power. After all, the sellers do contribute something; without them, the buyer could get nothing. If one could be sure that the sellers will form a cartel to bargain with the buyer, a reasonable compromise would be (1/2, 1/4, 1/4). In fact, the value is (2/3, 1/6, 1/6), representing something between the cartel solution and the competitive one; a cartel is possible, but is not a certainty.

Consider next a market in two perfectly divisible and completely complementary goods, which we may call right and left gloves. There are four players; initially 1 and 2 hold one and two left gloves respectively, 3 and 4 hold one right glove each. In coalitional form, $v(1234) = v(234) = 2$, $v(ij) = v(12j) = v(134) = 1$,$v(S) = 0$ otherwise, where $I = 1, 2$, and $j = 3,4$.The core consists of (0, 0, 1, 1) only; that is, the owners of the left gloves must simply give away their merchandise, for nothing. This in itself seems strange enough. It becomes even stranger when one realizes that Player 2 could make the situation entirely symmetric (as between 1, 2 and 3, 4) simply by burning one glove, an action that he can take alone, without consulting anybody.

The value can never suffer from this kind of pathological breakdown in monotonicity. Here $\varphi v = (1/4,7/12,7/12,7/12,)$ which nicely reflects the features of the situation. There *is* an oversupply of left gloves, and three and four do benefit from it. Also two benefits from it; he always has the option of nullifying it, but he can also use it (when he has an opportunity to strike a deal with both 3 and 4). The brunt of the oversupply is thus born by 1 who, unlike 2, cannot take measures to correct it.

Finally, consider a market with 2,000,001 players, 1,000,000 holding one right glove each, and 1,000,001 holding one left glove each. Again, the core stipulates that the holders of the left gloves must all give away their merchandise, for nothing. True, there *is* a slight oversupply of left gloves; but one would hardly have imagined so drastic an effect from one single glove out of millions. The value, too, takes the oversupply into account, but not in such an extreme form; altogether, the left-glove holders get about 499,557 pairs, the right about 500,443 (Shapley and Shubik 1969b). This is much more reasonable, though the effect is still surprisingly large: The short side gains an advantage that amounts to almost a thousand pairs.

The value has many different characterizations, all of them intuitively meaningful and interesting. We have already mentioned Shapley's original axioms, the value formula, and Young's characterization. To them must be added Harsanyi's (1959) dividend characterization, Owen's (1972) fuzzy coalition formula, Myerson's (1977) graph approach, Dubey's (1980) diagonal formula, the potential of Hart and Mas-Colell (1986), the reduced game axiomatization by the same authors, and Roth's (1977) formalization of Shapley's (1953b) idea that the value represents the utility to the players of playing a game. Moreover, because of its mathematical tractability, the value lends itself to a far greater range of applications than any other cooperative solution concept. And in terms of general theorems and characterizations for wide classes of games and economies, the value has a greater range than *any* other solution concept, barnone.

Previously (*1930–1950*, iii), we compared solution concepts of games to indicators of distributions, like mean and median. In fact the value is in many ways analogous to the mean, whereas the median corresponds to something like the core, or to core-like concepts such as the nucleolus (*1960–1970*, iv). Like the core, the median has an intuitively transparent and compelling definition (the point that cuts the distribution exactly in half), but lacks an algebraically neat formula; and like the value, the mean has a neat formula whose intuitive significance is not entirely transparent (thought through much experience from childhood on, many people have acquired an intuitive feel for it). Like the value, the mean is linear in its data; the core, nucleolus, and median are not. Both the mean and the value are very sensitive to their data: change one datum by a little, and the mean (or value) will respond in the appropriate direction; neither the median nor the core is sensitive in this way: one can change the data in wide ranges without affecting the median (or core) at all. On the other hand, the median can suddenly jump because of a moderate change in just one datum; thus the median of 1,000,001 zeros and 1,000,000 ones is 0, but jumps to 1 if we change just one datum from 0 to 1. We have already seen that the core may behave similarly, but the mean and the value cannot. Both the mean and the value are mathematically very tractable, resulting in a wide range of applications, both theoretical and practical; the median and core are less tractable, resulting in a narrower (though still considerable) range of applications.

The first extensive applications of the value were to various voting games (Shapley and Shubik 1954). The key observation in this seminal paper was that the value of a player equals his probability of *pivoting* – turning a coalition from losing to winning – when the players are ordered at random. From this there has grown a very large literature on voting games. Other important classes of applications are to market games (*1960–1970*, v) and political-economic games (e.g. Aumann and Kurz 1977; Neyman 1985b).

(vii) *Axiomatics*. The Shapley value and Nash's solution to the bargaining problem are examples of the axiomatic approach. Rather than defining a solution concept directly, one writes down a set of conditions for it to satisfy, then sees where they lead. In many contexts, even a relatively small set of fairly reasonable conditions turn out to be self-contradictory; there is no concept satisfying all of them. The most famous instance of this is Arrow's (1951) impossibility theorem for social welfare functions, which is one of the earliest applications of axiomatics in the social sciences.

It is not easy to pin down precisely what is meant by 'the axiomatic method'. Sometimes the term is used for any formal deductive system, with undefined terms, assumptions, and conclusions. As understood today, all of game theory and mathematical economics fits that definition. More narrowly construed, an axiom system is a small set of individually transparent conditions, set in a fairly general and abstract framework, which when taken together have far-reaching implications. Examples are Euclid's axioms for geometry, the Zermelo-Fraenkel axioms for set theory, the conditions on multiplication that define a group, the conditions on open sets that define a topological space, and the conditions on preferences that define utility and/or subjective probability.

Game theoretic solution concepts often have both direct and axiomatic characterizations. The direct definition applies to each game separately, whereas most axioms deal with relationships between games. Thus the formula for the Shapley value φv enables one to calculate it without referring to any game other than v. But the axioms for φ concern relationships between games; they say that if the values of certain games are so and so, then the values of certain other, related games must be such and such. For example, the additivity axiom is $\varphi(v + w) = \varphi v + \varphi w$. This is analogous to direct vs. axiomatic approaches to integration. Direct approaches such as limit of sum work on a single function; axiomatic approaches characterize the integral as a linear operator on a *space* of functions. (Harking back to the discussion at (vi), we note that the axioms for the value are quite similar to those for the

integral, which in turn is closely related to the mean of a distribution.)

Shapley's value and the solutions to the bargaining problem due to Nash (1950), Kalai and Smorodinsky (1975), and Maschler and Perles (1981) were originally conceived axiomatically, with the direct characterization coming afterwards. In other cases the process was reversed; for example, the nucleolus, NTU Shapley value, and NTU Harsanyi value were all axiomatized only years after their original direct definition (see *1960–1970)*. Recently the core, too, has been axiomatized (Peleg 1985, 1986).

Since axiomatizations concern relations between different games, one may ask why the players of a given game should be concerned with other games, which they are not playing. This has several answers. Viewed as an indicator, a solution of a game doesn't tell us much unless it stands in some kind of coherent relationship to the solutions of other games. The ratings for a rock climb tell you something if you have climbed other rocks whose ratings you know; topographic maps enable you to take in a situation at a glance if you have used them before, in different areas. If we view a solution as an arbitrated or imposed outcome, it is natural to expect some measure of consistency from an arbitrator or judge. Indeed, much of the law is based on precedent, which means relating the solution of the given 'game' to those of others with known solutions. Even when viewing a solution concept as a norm of actual behaviour, the very word 'norm' implies that we are thinking of a function on classes of games rather than of a single game; outcomes are largely based on mutual expectations, which are determined by previous experience with other games, by 'norms'.

Axiomatizations serve a number of useful purposes. First, like any other alternative characterization, they shed additional light on a concept, enable us to 'understand' it better. Second, they underscore and clarify important similarities between concepts, as well as differences between them. One example of this is the remarkable 'reduced game property' or 'consistency principle', which is associated in various different forms with just about every solution concept, and plays a key role in many of the axiomatizations (see *1970–1986*, vi). Another example consists of the axiomatizations of the Shapley and Harsanyi NTU values. Here the axioms are exact analogues, except that in the Shapley case they refer to payoff profiles, and in the Harsanyi case to 2^n-tuples of payoff profiles, one for each of the 2^n coalitions (Hart 1985a). This underscores the basic difference in outlook between those two concepts: The Shapley value assumes that the all-player coalition eventually forms, the intermediate coalitions being important only for bargaining chips and threats, whereas the Harsanyi value takes into account a real possibility of the intermediate coalitions actually forming.

Last, an important function of axiomatics relates to 'counter-intuitive examples', in which a solution concept yields outcomes that seem bizarre; e.g. the cores of some of the games discussed above in (vi). Most axioms appearing in axiomatizations do seem reasonable on the face of it, and many of them are in fact quite compelling. The fact that a relatively small selection of such axioms is often categoric (determines a unique solution concept), and that different such selections yield different answers, implies that all together, these reasonable-sounding axioms are contradictory. This, in turn, implies that any one solution concept will necessarily violate at least some of the axioms that are associated with other solution concepts; thus if the axioms are meant to represent intuition, counter-intuitive examples are inevitable.

In brief, axiomatics underscores the fact that a 'perfect' solution concept is an unattainable goal, a *fata morgana*; there is something 'wrong', some quirk with every one. Any given kind of counter-intuitive example can be eliminated by an appropriate choice of solution concept, but only at the cost of another quirk turning up. Different solution concepts can therefore be thought of as results of choosing not only which properties one likes, but also which examples one wishes to avoid.

1960–1970

The 1960s were a decade of growth. Extensions such as games of incomplete information and

NTU coalitional games made the theory much more widely applicable. The fundamental underlying concept of common knowledge was formulated and clarified. Core theory was extensively developed and applied to market economies; the bargaining set and related concepts such as the nucleolus were defined and investigated; games with many players were studied in depth. The discipline expanded geographically, outgrowing the confines of Princeton and Rand; important centres of research were established in Israel, Germany, Belgium and the Soviet Union. Perhaps most important was the forging of a strong, lasting relationship with mathematical economics and economic theory.

(i) *NTU coalitional games and NTU value*. Properly interpreted, the coalitional form (*1930–1950*, ii) applies both to TU and to NTU games; nevertheless, for many NTU applications one would like to describe the opportunities available to each coalition more faithfully than can be done with a single number. Accordingly, define a game in *NTU coalitional form* as a function that associates with each coalition S a set V(S) of S-tuples of real numbers (functions from S to $\mathbb{R}$). Intuitively, V(S) represents the set of payoff S-tuples that S can achieve. For example, in an exchange economy, V(S) is the set of utility S-tuples that S can achieve when its members trade among themselves only, without recourse to agents outside of S. Another example of an NTU coalitional game is Nash's bargaining problem (1950–1960, iii), where one can take V({1,2}) C,V(1) = {0},V(2) = {0}.

The definitions of stable set and core extend straightforwardly to NTU coalitional games, and these solution concepts were among the first to be investigated in that context (Aumann and Peleg 1960; Peleg 1963a; Aumann 1961). The first definitions of NTU value were proposed by Harsanyi (1959, 1963), but they proved difficult to apply. Building on Harsanyi's work, Shapley (1969a, b, c) defined a value for NTU games that has proved widely applicable and intuitively appealing.

For each profile λ of non-negative numbers and each outcome x, define the *weighted out come* λx by $(\lambda x)^i = \lambda^i x^i$. Let v_λ (S) be the maximum total weight that the coalition S can achieve,

$$v_\lambda(\mathrm{S}) := \max\left\{ \sum_{i \in \mathrm{S}} \lambda^i x^i, x \in \mathrm{V}(\mathrm{S}) \right\}.$$

Call an outcome x an *NTU value* of V if $x \in$ V (N) and there exists a weight profile λ with $\lambda x = \varphi v_\lambda$; in words, if x is feasible and corresponds to the value of one of the coalitional games v_λ.

Intuitively, v_λ(S) is a numerical measure of S's total worth and hence $\varphi^i v_\lambda$ measures i's social productivity. The weights λ^i are chosen so that the resulting value is feasible; an infeasible result would indicate that some people are overrated (or underrated), much like an imbalance between supply and demand indicates that some goods are overpriced (or underpriced).

The NTU value of a game need not be unique. This may at first sound strange, since unlike stability concepts such as the core, one might expect an 'index of social productivity' to be unique. But perhaps it is not so strange when one reflects that even a person's net worth depends on the prevailing (equilibrium) prices, which are not uniquely determined by the exogenous description of the economy.

The Shapley NTU value has been used in a very wide variety of economic and political-economic applications. To cite just one example, the Nash bargaining problem has a unique NTU value, which coincides with Nash's solution. For a partial bibliography of applications, see the references of Aumann (1985).

We have discussed the historical importance of TU as pointing the way for NTU results (*1930–1950*, vi). There is one piquant case in the reverse direction. Just as positive results are easier to obtain for TU, negative results are easier for NTU. Non-existence of stable sets was first discovered in NTU games (Stearns 1967), and this eventually led to Lucas's famous example (1969) of non-existence for TU.

(ii) *Incomplete information*. In 1957, Luce and Raiffa wrote that a fundamental assumption

G

> of game theory is that 'each player ... is fully aware of the rules of the game and the utility functions of each of the players ... this is a serious idealization which only rarely is met in actual situations' (p. 49). To deal with this problem, John Harsanyi (1967) constructed the theory of games of incomplete information (sometimes called differential or asymmetric information). This major conceptual breakthrough laid the theoretical groundwork for the great blooming of information economics that got under way soon thereafter, and that has become one of the major themes of modern economics and game theory.

For simplicity, we confine attention to strategic form games in which each player has a fixed, known set of strategies, and the only uncertainty is about the utility functions of the other players; these assumptions are removable. Bayesian rationality in the tradition of Savage (1954) dictates that all uncertainty can be made explicit; in particular, each player has a personal probability distribution on the possible utility (payoff) functions of the other player. But these distributions are not sufficient to describe the situation. It is not enough to specify what each player thinks about the other's payoffs; one must also know what he thinks they think about his (and each others') payoffs, what he thinks they think he thinks about their pay offs, and so on. This complicated infinite regress would appear to make useful analysis very difficult.

To cut this Gordian knot, Harsanyi postulated that each player may be one of several *types*, where a type determines both a player's own utility function and his personal probability distribution on the types of the other players. Each player is postulated to know his own type only. This enables him to calculate what he thinks the other players' types – and therefore their utilities – are. Moreover, his personal distribution on their types also enables him to calculate what he thinks they think about his type, and therefore about his utility. The reasoning extends indefinitely, and yields the infinite regress discussed above *as an outcome*.

Intuitively, one may think of a player's type as a possible state of mind, which would determine his utility as well as his distribution over others' states of mind. One need not assume that the number of states of mind (types) is finite; the theory works as well for, say, a continuum of types. But even with just two players and two types for each player, one gets a non-trivial infinite string of beliefs about utilities, beliefs about beliefs, and so on.

A model of this kind – with players, strategies, types, utilities, and personal probability distributions – is called an *I-game* (incomplete information game). A *strategic equilibrium* in an I-game consists of a strategy for each *type* of each player, which maximizes that type's expected payoff given the strategies of the other players' types.

Harsanyi's formulation of I-games is primarily a device for thinking about incomplete information in an orderly fashion, bringing that wild, bucking infinite regress under conceptual control. An (incomplete) analogy is to the strategic form of a game, a conceptual simplification without which it is unlikely that game theory would have gotten very far. Practically speaking, the strategic form of a particular game such as chess is totally unmanageable, one can't even begin to write it down. The advantage of the strategic form is that it is a comparatively simple formulation, mathematically much simpler than the extensive form; it enables one to formulate and calculate examples, which suggest principles that can be formulated and proved as general theorems. All this would be much more difficult – probably unachievable – with the extensive form; one would be unable to see the forest for the trees. A similar relationship holds between Harsanyi's I-game formulation and direct formulations in terms of beliefs about beliefs. (Compare the discussion of perspective made in connection with the coalitional form (*1930–1950*, i). That situation is somewhat different, though, since in going to the coalitional form, substantive information is lost. Harsanyi's formulation of I-games loses no information; it is a more abstract and simple – and hence transparent and workable – formulation of the same data as would be contained in an explicit description of the infinite regress.)

Harsanyi called an I-game *consistent* if all the personal probability distributions of all the types

are derivable as posteriors from a single prior distribution p on all n-tuples of types. Most applications of the theory have assumed consistency. A consistent I-game is closely related to the ordinary strategic game (*C-game*) obtained from it by allowing 'nature' to choose an n-tuple of types at random according to the distribution p, then informing each player of his type, and then playing the I-game as before. In particular, the strategic equilibria of a consistent I-game are essentially the same as the strategic equilibria of the related C-game. In the cooperative theory, however, an I-game is rather different from the related C-game, since binding agreements can only be made after the players know their types. Bargaining and other cooperative models have been treated in the incomplete information context by Harsanyi and Selten (1972), Wilson (1978), Myerson (1979, 1984), and others.

In a repeated game of incomplete information, the same game is played again and again, but the players do not have full information about it; for example, they may not know the others' utility functions. The actions of the players may implicitly reveal private information, e.g. about preferences; this may or may not be advantageous for them. We have seen (*1950–1960*, iii) that repetition may be viewed as a paradigm for cooperation. Strategic equilibria of repeated games of incomplete information may be interpreted as a subtle bargaining process, in which the players gradually reach wider and wider agreement, developing trust for each other while slowly revealing more and more information (Hart 1985b).

(iii) *Common knowledge*. Luce and Raiffa, in the statement quoted at the beginning of (ii), missed a subtle but important point. It is not enough that each player be fully aware of the rules of the game and the utility functions of the players. Each player must also be aware of this fact, i.e. of the awareness of all the players; moreover, each player must be aware that each player is aware that each player is aware, and so on ad infinitum. In brief, the awareness of the description of the game by all players must be a part of the description itself.

There is evidence that game theorists had been vaguely cognizant of the need for some such requirement ever since the late 1950s or early 1960s; but the first to give a clear, sharp formulation was the philosopher D.K. Lewis (1969). Lewis defined an event as *common knowledge* among a set of agents if all know it, all know that all know it, and so on adinfinitum.

The common knowledge assumption underlies all of game theory and much of economic theory. Whatever be the model under discussion, whether complete or incomplete information, consistent or inconsistent, repeated or one-shot, cooperative or non-cooperative, the model itself must be assumed common knowledge; otherwise the model is insufficiently specified, and the analysis incoherent.

(iv) *Bargaining set, kernel, nucleolus*. The core excludes the unique symmetric outcome (1/3, 1/3, 1/3) of the three-person voting game, because any two-person coalition can improve upon it. Stable sets (*1930–1950*, v) may be seen as a way of expressing our intuitive discomfort with this exclusion. Another way is the bargaining set (Davis and Maschler 1965). If, say, I suggests (1/2, 1/2, 0) to replace (1/3, 1/3, 1/3), then 3 can suggest to 2 that he is as good a partner as 1; indeed, 3 can even offer 2/3 to 2, still leaving himself with the 1/3 he was originally assigned. Formally, if we call (1/2, 1/2, 0) an *objection* to (1/3, 1/3, 1/3), then (0, 2/3, 1/3) is a *counter objection*, since it yields to 3 at least as much as he was originally assigned, and yields to 3's partners in the counter-objection at least as much as they were assigned either originally or in the objection. In brief, the counter-objecting player tells the objecting one, 'I can maintain my level of payoff and that of my partners, while matching your offers to players we both need.' An imputation is in the core if there is no objection to it. It is in the *bargaining set* if there is no *justified* objection to it, i.e. one that has no counter-objection.

Like the stable sets, the bargaining set includes the core (dominating and objecting are essentially

the same). Unlike the core and the set of stable sets, the bargaining set is for TU games never empty (Peleg 1967). For NTU it may be empty (Peleg 1963b); but Asscher (1976) has defined a non-empty variant; see also Billera (1970a).

Crucial parameters in calculating whether an imputation x is in the bargaining set of v are the *excesses* $v(S) - x(S)$ of coalitions S w.r.t. x, which measure the ability of members of S to use x in an objection (or counter-objection). Not, as is often wrongly assumed, because the initiator of the objection can assign the excess to himself while keeping his partners at their original level, but for precisely the opposite reason: because he can parcel out the excess to his partners, which makes counter objecting more difficult.

The excess is so ubiquitous in bargaining set calculations that it eventually took on intuitive significance on its own. This led to the formulation of two additional solution concepts: the *kernel* (Davis and Maschler 1965), which is always included in the bargaining set but is often much smaller, and the *nucleolus* (Schmeidler 1969), which always consists of a single point in the kernel.

To define the nucleolus, choose first all those imputations x whose maximum excess (among the 2^n excesses $v(S) = x(S)$) is minimum (among all imputations). Among the resulting imputations, choose next those whose second largest excess is minimum, and so on. Schmeidler's theorem asserts that by the time we have gone through this procedure 2^n times, there is just one imputation left.

We have seen that the excess is a measure of a coalition's 'manoeuvring ability'; in these terms the greatest measure of stability, as expressed by the nucleolus, is reached when all coalitions have manoeuvring ability as nearly a like as possible. An alternative interpretation of the excess is as a measure of S's total dissatisfaction with x, the volume of the cry that S might raise against x. In these terms, the nucleolus suggests that the final accommodation is determined by the loudest cry against it. Note that the *total* cry is determining, not the average cry; a large number of moderately unhappy citizens can be as potent a force for change as a moderate number of very unhappy ones. Variants of the nucleolus that use the average excess miss this point.

When the core is non-empty, the nucleolus is always in it. The nucleolus has been given several alternative characterizations, direct (Kohlberg 1971, 1972) as well as axiomatic (Sobolev 1975). The kernel was axiomatically characterized by Peleg (1986), and many interesting relationships have been found between the bargaining set, core, kernel, and nucleolus (e.g. Maschler et al. 1979). There is a large body of applications, of which we here cite just one: In a decisive weighted voting game, the nucleolus constitutes a set of weights (Peleg 1968). Thus the nucleolus may be thought of as a natural generalization of 'voting weights' to arbitrary games. (We have already seen that value and weights are quite different: see *1950–1960*, vi.)

(v) *The equivalence principle*. Perhaps the most remarkable single phenomenon in game and economic theory is the relationship between the price equilibria of a competitive market economy, and all but one of the major solution concepts for the corresponding game (the one exception is the stable set, about which more below). By a 'market economy' we here mean a pure exchange economy, or a production economy with constant returns.

We call an economy 'competitive' if it has many agents, each individual one of whom has too small an endowment to have a significant effect. This has been modelled by three approaches. In the *asymptotic approach*, one lets the number of agents tend to infinity, and shows that in an appropriate sense, the solution concept in question – core, value, bargaining set, or strategic equilibrium – tends to the set of competitive allocations (those corresponding to price equilibria). In the *continuum approach*, the agents constitute a (non-atomic) continuum, and one shows that the solution concept in question actually equals the set of competitive allocations (see the entry on large economies). In the *non-standard* approach, the agents constitute a non-standard model of the integers in the sense of Robinson (1974), and again one gets equality. Both the

continuum and the non-standard approaches require extensions of the theory to games with infinitely many players; see vi.

Intuitively, the equivalence principle says that the institution of market prices arises naturally from the basic forces at work in a market, (almost) no matter what we assume about the way in which these forces work. Compare (*1930–1950*, ix).

For simplicity in this section, unless otherwise indicated, the terms 'core', 'value', etc., refer to the limiting case. Thus 'core' means the limit of the cores of the finite economies, or the core of the continuum economy, or of the non-standard economy.

For the core, the asymptotic approach was pioneered by Edgeworth (1881), Shubik (1959a, b), and Debreu and Scarf (1963). Anderson (1986) is an excellent survey of the large literature that ensued. Early writers on the continuum approach included Aumann (1964) and Vind (1965); the non-standard approach was developed by Brown and Robinson (1975). Except for Shubik's, all these contributions were NTU. See the entry on CORE. After a 20-year courtship, this was the honeymoon of game theory and mathematical economics, and it is difficult to convey the palpable excitement of those early years of intimacy between the two disciplines.

Some early references for the value equivalence principle, covering both the asymptotic and continuum approaches, were listed above (*1930–1950*, vi). For the non-standard approach, see Brown and Loeb (1976). Whereas the core of a competitive economy equals the set of *all* competitive allocations, this holds for the value only when preferences are smooth (Shapley 1964; Aumann and Shapley 1974; Aumann 1975; Mas-Colell 1977). Without smoothness, every value allocation is competitive, but not every competitive allocation need be a value allocation. When preferences are kinky (non-differentiable utilities), the core is often quite large, and then the value is usually a very small subset of the core; it gives much more information. In the TU case, for example, the value is always a single point, even when the core is very large. Moreover, it occupies a central position in the core (Hart 1980; Tauman 1981; Mertens 1988); in particular, when the core has a centre of symmetry, the value is that centre of symmetry (Hart 1977a).

For example, suppose that in a glove market (*1950–1960*, vi), the number (or measure) of left-glove holders equals that of right-glove holders. Then at a price equilibrium, the price ratio between left and right gloves may be anything between 0 and ∞ (inclusive!). Thus the left-glove holders may end up giving away their merchandise for nothing to the right-glove holders, or the other way around, or anything in between. The same, of course, holds for the core. But the value prescribes precisely equal prices for right and left gloves.

It should be noted that in a finite market, the core contains the competitive allocations, but usually also much more. As the number of agents increases, the core 'shrinks', in the limit leaving only the competitive allocations. This is not so for the value; in finite markets, the value allocations may be disjoint from the core, and a fortiori from the competitive allocations (*1950–1960*, vi).

We have seen (*1930–1950*, iv) that the core represents a very strong and indeed not quite reasonable notion of stability. It might therefore seem perhaps not so terribly surprising that it shrinks to the competitive allocations. What happens, one may ask, when one considers one of the more reasonable stability concepts that are based on domination, such as the bargaining set or the stable sets?

For the bargaining set of TU markets, an asymptotic equivalence theorem was established by Shapley and Shubik in the mid-70s, though it was not published until 1984. Extending this result to NTU, to the continuum, or to both seemed difficult. The problems were conceptual as well as mathematical; it was difficult to give a coherent formulation. In 1986, Shapley presented the TU proof at a conference on the equivalence principle that took place at Stony Brook. A. Mas-Colell, who was in the audience, recognized the relevance of several results that he had obtained in other connections; within a day or 2 he was able to formulate and prove the equivalence principle for the bargaining set in NTU continuum economies (Mas-Colell 1988). In particular, this implies the

core equivalence principle; but it is a much stronger and more satisfying result.

For the strategic equilibrium the situation had long been less satisfactory, though there were results (Shubik 1973; Dubey and Shapley 1980). The difficulty was in constructing a satisfactory strategic (or extensive) model of exchange. Very recently Douglas Gale (1986) provided such a model and used it to prove a remarkable equivalence theorem for strategic equilibria in the continuum mode.

The one notable exception to the equivalence principle is the case of stable sets, which predict the formation of cartels even in fully competitive economies (Hart 1974). For example, suppose half the agents in a continuum initially hold 2 units of bread each, half initially hold 2 units of cheese, and the utility functions are concave, differentiable, and symmetric (e.g., $u(x, y) = \sqrt{x} + \sqrt{y}$). There is then a unique price equilibrium, with equal prices for bread and cheese. Thus each agent ends up with one piece of bread and one piece of cheese; this is also the unique point in the core and in the bargaining set, and the unique NTU value. But stable set theory predicts that the cheese holders will form a cartel, the bread holders will form a cartel, and these two cartels will bargain with each other as if they were individuals. The upshot will depend on the bargaining, and may yield an outcome that is much better for one side than for the other. Thus at each point of the unique stable set with the full symmetry of the game, each agent on each side gets as much as each other agent on that side; but these two amounts depend on the bargaining, and may be quite different from each other.

In a sense, the failure of stable set theory to fall into line makes the other results even more impressive. It shows that there isn't some implicit tautology lurking in the background, that the equivalence principle makes a substantive assertion.

In the *Theory of Games*, von Neumann and Morgenstern (1944) wrote that

> when the number of participants becomes really great, some hope emerges that the influence of every particular participant will become negligible. These are, of course, the classical conditions of 'free competition' ... The current assertions concerning free competition appear to be very valuable surmises and inspiring anticipations of results. But they are not results, and it is scientifically unsound to treat them as such.

One may take the theorems constituting the equivalence principle as embodying precisely this kind of 'result'. Yet it is interesting that Morgenstern himself, who died in 1977, never became convinced of the validity of the equivalence principle; he thought of it as mathematically correct but economically wrongheaded. It was his firm opinion that economic agents organize themselves into coalitions, that perfect competition is a fiction, and that stable sets explain it all. The greatness of the man is attested to by the fact that though scientifically opposed to the equivalence principle, he gave generous support, both financial and moral, to workers in this area.

(vi) *Many players*. The preface to *Contributions to the Theory of Games* I (Kuhn and Tucker 1950) contains an agenda for future research that is remarkable in that so many of its items – computation of minimax, existence of stable sets, *n*-person value, NTU games, dynamic games – did in fact become central in subsequent work. Item 11 in this agenda reads, 'establish significant asymptotic properties of *n*-person games, for large *n*'. We have seen ((v)) how this was realized in the equivalence principle for large economies. But actually, political game models with many players are at least as old as economic ones, and may be older. During the early 1960s, L.S. Shapley, working alone and with various collaborators, wrote a series of seven memoranda at the Rand Corporation under the generic title 'Values of Large Games', several of which explored models of large elections, using the asymptotic and the continuum approaches. Among these were models which had both 'atoms' – players who are significant as individuals – and an 'ocean' of individually insignificant players. On example of this is a corporation with many small stockholders and a few large stockholders; see also Milnor and Shapley

(1978). 'Mixed' models of this kind – i.e. with an ocean as well as atoms – have been explored in economic as well as political contexts using various solution notions, and a large literature has developed. The core of mixed markets has been studied by Drèze et al. (1969), Gabszewicz and Mertens (1971), Shitovitz (1973) and many others. For the nucleolus of 'mixed' voting games, see Galil (1974). Among the studies of values of mixed games are Hart (1973), Fogelman and Quinzii (1980), and Neyman (1987).

Large games in which *all* the players are individually insignificant – *non-atomic* games – have also been studied extensively. Among the early contributions to value theory in this connection are Kannai (1966), Riker and Shapley (1968), and Aumann and Shapley (1974). The subject has proliferated greatly, with well over a hundred contributions since 1974, including theoretical contributions as well as economic and political applications.

There are also games with infinitely many players in which *all* the players are atoms, namely games with a denumerable infinity of players. Again, values and voting games loom large in this literature. See, e.g., Shapley (1962), Artstein (1972), and Berbee (1981).

(vii) *Cores of finite games and markets*. Though the core was defined as an independent solution concept by Gillies and Shapley already in the early 1950s, it was not until the 1960s that a significant body of theory was developed around it. The major developments centre around conditions for the core to be non-empty; gradually it came to be realized that such conditions hold most naturally and fully when the game has an 'economic' rather than a 'political' flavour, when it may be thought of as arising from a market economy.

The landmark contributions in this area were the following: the Gale and Shapley 1962 paper on the core of a marriage market; the work of Bondareva (1963) and Shapley (1967) on the balancedness condition for the non-emptiness of the core of a TU game; Scarf's 1967 work on balancedness in NTU games; the work of Shapley and Shubik (1969a) characterizing TU market games in terms of non-emptiness of the core; and subsequent work, mainly associated with the names of Billera and Bixby (1973), that extended the Shapley-Shubik condition to NTU games. Each of these contributions was truly seminal, in that it inspired a large body of subsequent work.

Gale and Shapley (1962) asked whether it is possible to match m women with m men so that there is no pair consisting of an unmatched woman and man who prefer each other to the partners with whom they were matched. The corresponding question for homosexuals has a negative answer: the preferences of four homosexuals may be such that no matter how they are paired off, there is always an unmatched pair of people who prefer each other to the person with whom they were matched. This is so, for example, if the preferences of a, b, and c are cyclic, whereas d is lowest in all the others' scales. But for the heterosexual problem, Gale and Shapley showed that the answer is positive.

This may be stated by saying that the appropriately defined NTU coalitional game has a non-empty core. Gale and Shapley proved not only the non-emptiness but also provided a simple algorithm for finding a point in it.

This work has spawned a large literature on the cores of discrete market games. One fairly general recent result is Kaneko and Wooders (1982), but there are many others. A fascinating application to the assignment of interns to hospitals has been documented by Roth (1984). It turns out that American hospitals, after 50 years of turmoil, finally developed in 1950 a method of assignment that is precisely a point in the core.

We come now to general conditions for the core to be non-empty. Call a TU game v *super additive at* a coalition U if $v(\mathrm{U}) \geq \sum_j v(\mathrm{S}_j)$ for any partition of U into disjoint coalition Sj. This may be strengthened by allowing partitions of U into disjoint 'part-time' coalitions θS, interpreted as coalitions S operating during a proportion θ of the time $0 \leq \theta \leq 1$. Such a partition is therefore a family $\{\theta_j \mathrm{S}_j\}$, where the total amount of time that

each player in U is employed is exactly 1; i.e., where $\Sigma_j \theta_j \chi_{Sj} = \chi_U$, where χ_S is the indicator function of S. If we think of v(S) as there venue that Scan generate when operating full-time, then the part-time coalition θS generates θv(S). Superadditivity at U for part-time coalitions thus means that

$$\sum_j \theta_j \chi_{S_j} = \chi_U \text{ implies } v(\mathrm{U}) \geq \sum_j \theta_j v(\mathrm{S}_j)$$

A TU game v obeying this condition for U = I is called *balanced*; for all U, *totally balanced*.

Intuitively, it is obvious that a game with a non-empty core must be superadditive at I; and once we have the notion of part-time coalitions, it is only slightly less obvious that it must be balanced. The converse was established (independently) by Bondareva (1963) and Shapley (1967). Thus *a TU game has a non-empty core if and only if it is balanced*.

The connection between the core and balancedness (generalized superadditivity) led to several lines of research. Scarf (1967) extended the notion of balancedness to NTU games, then showed that every balanced NTU game has a non-empty core. Unlike the Bondareva-Shapley proof, which is based on linear programming methods, Scarf's proof was more closely related to fixed-point ideas. Eventually, Scarf realized that his methods could be used actually to prove Brouwer's fixed-point theorem, and moreover, to develop effective algorithms for approximating fixed points. This, in turn, led to the development of algorithms for approximating competitive equilibria of economies (Scarf 1973), and to a whole area of numerical analysis dealing with the approximation of fixed points (see computation of general equilibria).

An extension of the Bondareva-Shapley result to the NTU case that is different from Scarf's was obtained by Billera (1970a).

Another line of research that grew out of balancedness deals with characterizing markets in purely game-theoretic terms. When can a given coalitional game v be expressed as a market game (*1930–1950*, ii)? The Bondareva-Shapley theorem implies that market games have non-empty cores, and this also follows from the fact that outcomes corresponding to competitive equilibria are always in the core. Since a subgame of a market game is itself a market game, it follows that *for v to be a market game, it is necessary that it and all its subgames have non-empty cores*, i.e., that the game be totally balanced. (A *subgame* of a coalitional game v is defined by restricting its domain to subcoalitions of a given coalition U.) Shapley and Shubik (1969a) showed that *this necessary condition is also sufficient*. Balancedness itself is not sufficient, since there exist games with non-empty cores having subgames with empty cores (e.g., |I| = 4, v(S): = 0, 0, 1, 1, 2 when |S| = 0, 1, 2, 3, 4, respectively).

For the NTU case, characterizations of market games have been obtained by Billera and Bixby (1973), Mas-Colell (1975), and others.

Though the subject of this section is finite markets, it is nevertheless worthwhile to relate the results to non-atomic games (where the players constitute a non-atomic continuum, an 'ocean'). The total balancedness condition then takes on a particularly simple form. Suppose, for simplicity, that v is a function of finitely many measures, i.e., $v(S) = f(\mu(S))$, where $\mu = (\mu_1,..., \mu_n)$, and the μ_j are non-atomic measures. Then v is a market game iff f is concave and one-homogeneous ($f(\theta x) = \theta(fx)$ when $\theta \geq 0$). This is equivalent to saying that v is superadditive (at all coalitions), and f is 1-homogeneous (Aumann and Shapley 1974).

Perhaps the most remarkable expression of the connection between superadditivity and the core has been obtained by Wooders (1983). Consider coalitional games with a fixed finite number k of 'types' of players, the coalitional form being given by $v(S) = f(\mu(S))$, where $\mu(S)$ is the profile of type sizes in S, i.e. it is a vector whose i'th coordinate represents the number of type i players in S. (To specify the game, μ(I) must also be specified.) Assume that f is superadditive, i.e. $f(x + y) \geq f(x) + f(y)$ for all x and y with non-negative integer coordinates; this assures the superadditivity of v. Moreover, assume that f obeys a 'Lipschitz' condition, namely that $|f(x) - f(y)|/\|x - y\|$ is uniformly bounded for all $x \neq y$, where $\|x\| : = max_j |x_j|$. Then for each

$\varepsilon > 0$, when the number of players is sufficiently large, the ε-core is non-empty. (The ε-core is defined as the set of all outcomes x such that $x(S) \geq v(S) - \varepsilon|S|$ for all S.) Roughly, the result says that the core is 'almost' non-empty for sufficiently large games that are superadditive and obey the Lipschitz condition. Intuitively, the superadditivity together with the Lipschitz condition yield 'approximate' 1-homogeneity, and in the presence of 1- homogeneity, superadditivity is equivalent to concavity. Thus f is approximately a 1- homogeneous concave function, so that we are back in a situation similar to that treated in the previous paragraph. What makes this result so remarkable is that other than the Lipschitz condition, the only substantive assumption is superadditivity.

Wooders (1983) also obtained a similar theorem for NTU; Wooders and Zame (1984) obtained a formulation that does away with the finite type assumption.

1970–1986

We do not yet have sufficient distance to see the developments of this period in proper perspective. Political and political economic models were studied in depth. Non-cooperative game theory was applied to a large variety of particular economic models, and this led to the study of important variants on the refinements of the equilibrium concept. Great strides forward were made in almost all the areas that had been initiated in previous decades, such as repeated games (both of complete and of incomplete information), stochastic games, value, core, nucleolus, bargaining theory, games with many players, and so on (many of these developments have been mentioned above). Game Theory was applied to biology, computer science, moral philosophy, cost allocation. New light was shed on old concepts such as randomized strategies.

Sociologically, the discipline proliferated greatly. Some 16 or 17 people participated in the first international workshop on game theory held in Jerusalem in 1965; the fourth one, held in Cornell in 1978, attracted close to 100, and the discipline is now too large to make such workshops useful. An international workshop in the relatively restricted area of repeated games, held in Jerusalem in 1985, attracted over 50 participants. The *International Journal of Game Theory* was founded in 1972; *Mathematics of Operations Research*, founded in 1975, was organized into three major 'areas', one of them Game Theory. Economic theory journals, such as the *Journal of Mathematical Economics*, the *Journal of Economic Theory, Econometrica*, and others devoted increasing proportions of their space to game theory. Important centres of research, in addition to the existing ones, sprang up in France, Holland, Japan, England, and India, and at many Universities in the United States.

Gradually, game theory also became less personal, less the exclusive concern of a small 'in' group whose members all know each other. For years, it had been a tradition in game theory to publish only a fraction of what one had found, and then only after great delays, and not always what is most important. Many results were passed on by word of mouth, or remained hidden in ill-circulated research memoranda. The 'Folk Theorem' to which we alluded above (*1950–1960*, iii) is an example. This tradition had both beneficial and deleterious effects. On the one hand, people did not rush into print with trivia, and the slow cooking of results improved their flavour. As a result, phenomena were sometimes rediscovered several times, which is perhaps not entirely bad, since you understand something best when you discover it yourself. On the other hand, it was difficult for outsiders to break in; non-publication caused less interest to be generated than would otherwise have been, and significantly impeded progress.

Be that as it may, those days are over. There are now hundreds of practitioners, they do not all know each other, and sometimes have never even heard of one another. It is no longer possible to communicate in the old way, and as a result, people are publishing more quickly. As in other disciplines, it is becoming difficult to keep abreast of the important developments. Game theory has matured.

(i) *Applications to biology.* A development of outstanding importance, whose implications are not yet fully appreciated, is the application of game theory to evolutionary *biology*. The high priest of this subject is John Maynard Smith (1982), a biologist whose concept of *evolutionarily stable strategy*, a variant of strategic equilibrium, caught the imagination both of biologists and of game theorists. On the game theoretic side, the theme was taken up by Reinhard Selten (1980, 1983) and his school; a conference on 'Evolutionary theory in biology and economics', organized by Selten in Bielefeld in 1985, was enormously successful in bringing field biologists together with theorists of games to discuss these issues. A typical paper was tit for tat in the great tit (Regelmann and Curio 1986); using actual field observations, complete with photographs, it describes how the celebrated 'tit for tat' strategy in the repeated prisoners' dilemma (Axelrod 1984) accurately describes the behaviour of males and females of a rather common species of bird called the great tit, when protecting their young from predators.

It turns out that ordinary, utility maximizing rationality is much more easily observed in animals and even plants than it is in human beings. There are even situations where rats do significantly better than human beings. Consider, for example, the famous probability matching experiment, where the subject must predict the values of a sequence of i.i.d. random variables taking the values L and R with probabilities 3/4 and 1/4 respectively; each correct prediction is rewarded. It is of course optimal always to predict L; but human subjects tend to match the probabilities, i.e. to predict L about 3/4 of the time. On the other hand, while rats are not perfect (i.e. do not predict L *all* the time), they do predict L significantly more often than human beings.

Several explanations have been suggested. One is that in human experimentation, the subjects try subconsciously to 'guess right', i.e. to guess what the experimenter 'wants' them to do, rather than maximizing utility. Another is simply that the rats are more highly motivated. They are brought down to 80% of their normal body weight, are literally starving; it is much more important for them to behave optimally than it is for human subjects.

Returning to theory, though the notion of strategic equilibrium seems on the face of it simple and natural enough, a careful examination of the definition leads to some doubts and questions as to why and under what conditions the players in a game might be expected to play a strategic equilibrium. See the entry on Nash equilibrium, refinements of. Evolutionary theory suggests a simple rationale for strategic equilibrium, in which there is no conscious or overt decision making at all. For definiteness, we confine attention to two-person games, though the same ideas apply to the general case. We think of each of the two players as a whole species rather than an individual; reproduction is assumed asexual. The set of pure strategies of each player is interpreted as the locus of some gene (examples of a locus are eye colour, degree of aggressiveness, etc.); individual pure strategies are interpreted as alleles (blue or green or brown eyes, aggressive or timid behaviour, etc.). A given individual of each species possesses just one allele at the given locus; he interacts with precisely one individual in the other species, who also has just one allele at the locus of interest. The result of the interaction is a definite increment or decrement in the fitness of each of the two individuals, i.e. the number (or expected number) of his offspring; thus the payoff in the game is denominated in terms of fitness.

In these terms, a mixed strategy is a distribution of alleles throughout the population of the species (e.g., 40% aggressive, 60% timid). If each individual of each species is just as likely to meet any one individual of the other species as any other one, then the probability distribution of alleles that each individual faces is precisely given by the original mixed strategy. It then follows that a given pair of mixed strategies is a strategic equilibrium if and only if it represents a population equilibrium, i.e. a pair of distributions of characteristics (alleles) that does not tend to change.

Unfortunately, sexual reproduction screws up this story, and indeed the entire Maynard Smith approach has been criticized for this reason. But to be useful, the story does not have to be taken entirely literally. For example, it applies to evolution that is cultural rather than biological. In this approach, a 'game' is interpreted as a *kind* of confrontational situation (like shopping for a car) rather than a specific instance of such a situation; a 'player' is a role ('buyer' or 'salesman'), not an individual human being; a pure strategy is a possible kind of behaviour in this role ('hard sell' or 'soft sell'). Up to now this is indeed not very different from traditional game theoretic usage. What is different in the evolutionary interpretation is that pure or mixed strategic equilibria do not represent conscious rational choices of the players, but rather a population equilibrium which evolves as the result of how successful certain behaviour is in certain roles.

(ii) *Randomization as ignorance*. In the traditional view of strategy randomization, the players use a randomizing device, such as a coin flip, to decide on their actions. This view has always had difficulties. Practically speaking, the idea that serious people would base important decisions on the flip of a coin is difficult to swallow. Conceptually, too, there are problems. The reason a player must randomize in equilibrium is only to keep others from deviating. For himself, randomizing is unnecessary; he will do as well by choosing any pure strategy that appears with positive probability in his equilibrium mixed strategy.

Of course, there is no problem if we adopt the evolutionary model described above in (i); mixed strategies appear as population distributions, and there is no explicit randomization at all. But what is one to make of randomization within the more usual paradigm of conscious, rational choice?

According to Savage (1954), randomness is not physical, but represents the ignorance of the decision maker. You associate a probability with every event about which you are ignorant, whether this event is a coin flip or a strategic choice by another player. The important thing in strategy randomization is that the *other* players be ignorant of what you are doing, and that they ascribe the appropriate probabilities to each of your pure strategies. It is not necessary for you actually to flip a coin.

The first to break away from the idea of explicit randomization was J. Harsanyi (1973). He showed that if the payoffs to each player i in a game are subjected to small independent random perturbations, known to i but not to the other players, then the resulting game of incomplete information has *pure* strategy equilibria that correspond to the mixed strategy equilibria of the original game. In plain words, nobody really randomizes. The appearance of randomization is due to the payoffs not being exactly known to all; each player, who does know his own payoff exactly, has a unique optimal action against his estimate of what the others will do.

This reasoning may be taken one step further. Even without perturbed payoffs, the players simply do not know which strategies will be chosen by the other players. At an equilibrium of 'matching pennies', each player knows very well what he himself will do, but ascribes 1/2 – 1/2 probabilities to the other's actions; he also knows that the other ascribes those probabilities to his own actions, though it is admittedly not quite obvious that this is necessarily the case. In the case of a general n-person game, the situation is essentially similar; the mixed strategies of i canal ways be understood as describing the uncertainty of players other than i about what i will do (Aumann 1987).

(iii) *Refinements of strategic equilibrium*. In analysing specific economic models using the strategic equilibrium – an activity carried forward with great vigour since about 1975 – it was found that Nash's definition does not provide adequately for rational choices given one's information at each stage of an extensive game. Very roughly, the reason is that Nash's definition ignores contingencies 'off the equilibrium path'. To remedy this, various "refinements' of strategic equilibrium have been defined, starting with Selten's (1975) 'trembling hand' equilibrium. Please

refer to our discussion of Zermelo's theorem (1930–1950, vi), and to Section IV of the entry on nash equilibrium, refinements of.

The interesting aspect of these refinements is that they use *irrationality* to arrive at a strong form of rationality. In one way or another, all of them work by assuming that irrationality cannot be ruled out, that the players ascribe irrationality to each other with a small probability. True rationality requires 'noise'; it cannot grow in sterile ground, it cannot feed on itself only.

(iv) *Bounded rationality.* For a long time it has been felt that both game and economic theory assume too much rationality. For example the hundred-times repeated prisoner's dilemma has some $2^{2^{100}}$ pure strategies; all the books in the world are not large enough to write this number even once in decimal notation. There is no practical way in which all these strategies can be considered truly available to the players. On the face of it, this would seem to render statements about the equilibrium points of such games (*1950–1960*, iv) less compelling, since it is quite possible that if the sets of strategies were suitably restricted, the equilibria would change drastically.

For many years, little on the formal level was done about these problems. Recently the theory of automata has been used for formulations of bounded rationality in repeated games. Neyman (1985a) assumes that only strategies that are programmable on an automaton of exogenously fixed size can be considered 'available' to the players. He then shows that even when the size is very large, one obtains results that are qualitatively different from those when all strategies are permitted. Thus in the n-times repeated prisoner's dilemma, only the greedy-greedy outcome can occur in equilibrium; but if one restricts the players to using automata with as many as $e^{o(n)}$ states, then for sufficiently large n, one can approximate in equilibrium any feasible individually rational outcome, and in particular the friendly–friendly outcome. For example, this is the case if the number of states is bounded by any fixed polynomial in n. In unpublished work, Neyman has generalized this result from the prisoner's dilemma to arbitrary games; specifically, he shows that a result similar to the Folk Theorem holds in any long finitely repeated game, when the automaton size is limited as above to subexponential.

Another approach has been used by Rubinstein (1986), with dramatically different results. In this work, the automaton itself is endogenous; all states of the automaton must actually be used on the equilibrium path. Applied to the prisoner's dilemma, this assumption leads to the conclusion that in equilibrium, one cannot get anywhere near the friendly–friendly outcome. Intuitively, the requirement that all states be used in equilibrium rules out strategies that punish deviations from equilibrium, and these are essential to the implicit enforcement mechanism that underlies the folk theorem. See the discussion at (*1950–1960*, iii) above.

(v) *Distributed computing*. In the previous subsection (iv) we discussed applications of computer science to game theory. There are also applications in the opposite direction; with the advent of distributed computing, game theory has become of interest in computer science. Different units of a distributed computing system are viewed as different players, who must communicate and coordinate. Breakdowns and failures of one unit are often modelled as malevolent, so as to get an idea as to how bad the worst case can be. From the point of view of computer tampering and crime, the model of the malevolent player is not merely a fiction; similar remarks hold for cryptography, where the system must be made proof against purposeful attempts to 'break in'. Finally, multi-user systems come close to being games in the ordinary sense of the word.

(vi) *Consistency* is a remarkable property which, in one form or another, is common to just about all game-theoretic solution concepts. Let us be given a game, which for definiteness we denote v, though it may be NTU or

even non-cooperative. Let x be an outcome that 'solves' the game in some sense, like the value or nucleolus or a point in the core. Suppose now that some coalition S wishes to view the situation as if the players outside S get their components of x so to speak exogenously, without participating in the play. That means that the players in S are playing the 'reduced game' v_x^S, whose all-player set is S. It is not always easy to say just how v_x^S should be defined, but let's leave that aside for the moment. Suppose we apply to v_x^S the same solution concept that when applied to v yields x. Then the consistency property is that $x|S$ (x restricted to S) is the resulting solution. For example, if x is the nucleolus of v, then for each v, the restriction $x|S$ is the nucleolus of v_x^S.

Consistency implies that it is not too important how the player set is chosen. One can confine attention to a 'small world', and the outcome for the denizens of this world will be the same as if we had looked at them in a 'big world'.

In a game theoretic context, consistency was first noticed by J. Harsanyi (1959) for the Nash solution to the n-person bargaining game. This is simply an NTU game V in which the only significant coalitions are the single players and the all-player coalition, and the single players are normalized to get 0. The Nash solution, axiomatized by Harsanyi (1959), is the outcome x that maximizes the product $x^1x^2 \ldots x^n$. To explain the consistency condition, let us look at the case $n = 3$, in which case $V(\{1, 2, 3\})$ is a subset of three-space. If we let $S = \{1, 2\}$, and if x_0 is the Nash solution, then 3 should get x_0^3. That means that 1 and 2 are confined to bargaining within that slice of $V\{(1, 2, 3)\}$ that is determined by the plane $x^3 = x_0^3$. According to the Nash solution for the two-person case, they should maximize x^1x^2 over this slice; it is not difficult to see that this maximum is attained at $\left(x_0^1, x_0^2\right)$, which is exactly what consistency requires.

Davis and Maschler (1965) proved that the kernel satisfies a consistency condition; so do the bargaining set, core, stable set, and nucleolus, using the same definition of the reduced game v_x^S as for the kernel (Aumann and Drèze 1974). Using a somewhat different definition of v_x^S, consistency can be established for the value (Hart and Mas-Colell 1986). Note that strategic equilibria, too, are consistent; if the players outside S play their equilibrium strategies, an equilibrium of the resulting game on S is given by having the players in S play the same strategies that they were playing in the equilibrium of the large game.

Consistency often plays a key role in axiomatizations. Strategic equilibrium is axiomatized by consistency, together with the requirement that in one-person maximization problems, the maximum be chosen. A remarkable axiomatization of the Nash solution to the bargaining problem (including the two-person case discussed at *1950–1960*, v), in which the key role is played by consistency, has been provided by T. Lensberg (1981). Axiomatizations in which consistency plays the key role have been provided for the nucleolus (Sobolev 1975), core (Peleg 1985, 1986), kernel (Peleg 1986), and value (Hart and Mas-Colell 1986). Consistency-like conditions have also been used in contexts that are not strictly game-theoretic, e.g. by Balinski and Young (1982), W. Thomson, J. Roemer, H. Moulin, H.P. Young and others.

In law, the consistency criterion goes back at least to the 2000-year old Babylonian Talmud (Aumann and Maschler 1985). Though it is indeed a very natural condition, its huge scope is still somewhat startling.

(vii) The fascination of *cost allocation* is that it retains the formal structure of cooperative game theory in a totally different interpretation. The question is how to allocate joint costs among users. For example, the cost of a water supply or sewage disposal system serving several municipalities (e.g. Bogardi and Szidarovsky 1976); the cost of telephone calls in an organization such as a university or corporation (Billera et al. 1978); or the cost of an airport (Littlechild and Owen 1973, 1976). In the airport case, for example, each 'player' is one landing of one airplane, and $v(S)$ is the cost of building and running an airport large enough to

accommodate the set S of landings. Note that $v(S)$ depends not only on the number of landings in S but also on its composition; one would not charge the same for a landing of a 747 as for a Piper, for example because the 747 requires a longer runway. The allocation of cost would depend on the solution concept; for example, if we are using the Shapley value φ, then the fee for each landing i would be $\varphi^i v$.

The axiomatic method is particularly attractive here, since in this application the axioms often have rather transparent meaning. Most frequently used has been the Shapley value, whose axiomatic characterization (see Shapley value) is particularly transparent (Billera and Heath 1982).

The literature on the game theoretic approach to cost allocation is quite large, probably several hundred items, many of them in the accounting literature (e.g. Roth and Verrecchia 1979).

Concluding Remarks

(i) *Ethics*. While game theory does have intellectual ties to ethics, it is important to realize that in itself, it has no moral content, makes no moral recommendations, is ethically neutral. Strategic equilibrium does not tell us to maximize utility, it explores what happens when we do. The Shapley value does not recommend dividing payoff according to power, it simply measures the power. Game Theory is a tool for telling us where incentives will lead. History and experience teach us that if we want to achieve certain goals, including moral and ethical ones, we had better see to the incentive effects of what we are doing; and if we do not want people to usurp power for themselves, we had better build institutions that spread power as thinly and evenly as possible. Blaming game theory – or, for that matter, economic theory – for selfishness is like blaming bacteriology for disease. Game theory studies selfishness, it does not recommend it.

(ii) *Mathematical methods*. We have had very little to say about mathematical methods in the foregoing, because we wished to stress the conceptual side. Worth noting, though, is that mathematically, game theoretic results developed in on context often have important implications in completely different contexts. We have already mentioned the implications of two-person zero-sum theory for the theory of the core and for correlated equilibria (*1910–1930*, vii). The first proofs of the existence of competitive equilibrium (Arrow and Debreu 1954) used the existence of strategic equilibrium in a generalized game (Debreu 1952). Blackwell's 1956 theory of two-person zero-sum games with vector payoffs is of fundamental importance for n-person repeated games of complete information (Aumann 1961) and for repeated games of incomplete information (e.g. Mertens 1982; Hart 1985b). The Lemke-Howson algorithm (1962) for finding equilibria of two-person non-zero sum non-cooperative games was seminal in the development of the algorithms of Scarf (1967, 1973) for finding points in the core and finding economic equilibria.

(iii) *Terminology*. Game theory has sometimes been plagued by haphazard, inappropriate terminology. Some workers, notably L.S. Shapley (1973b), have tried to introduce more appropriate terminology, and we have here followed their lead. What follows is a brief glossary to aid the reader in making the proper associations.

Used here	Older term
Strategic form	Normal form
Strategic equilibrium	Nash equilibrium
Coalitional form	Characteristic function
Transferable utility	Side payment
Decisive voting game	Strong voting game
Improve upon	Block
Worth	Characteristic function value
Profile	n-tuple
1-homogeneous	Homogeneous of degree 1

See Also

- ▶ Exchange
- ▶ Shapley Value

Bibliography

Anderson, R.M 1986. Notions of core convergence. In Hildebrand and Mas-Colell (1986), 25–46.

Arrow, K.J. 1951. *Social choice and individual values*. New York: Wiley.

Arrow, K.J., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Artstein, Z. 1972. Values of games with denumerable many players. *International Journal of Games Theory* 3: 129–140.

Asscher, N. 1976. An ordinal bargaining set for games without side payments. *Mathematics of Operations Research* 1: 381–389.

Aumann, R.J. 1960. Linearity of unrestrictedly transferable utilities. *Naval Research Logistics Quarterly* 7: 281–284.

Aumann, R.J. 1961. The core of a cooperative game without sidepayments. *Transactions of the American Mathematical Society* 98: 539–552.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.

Aumann, R.J. 1975. Values of markets with a continuum of traders. *Econometrica* 43: 611–646.

Aumann, R.J. 1985. On the non-transferable utility value: A comment on the Roth-Shafer examples. *Econometrica* 53: 667–677.

Aumann, R.J. 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55: 1–18.

Aumann, R.J., and J.H. Drèze. 1974. Cooperative games with coalition structures. *International Journal of Games Theory* 3: 217–238.

Aumann, R.J., and M. Kurz. 1977. Power and taxes. *Econometrica* 45: 1137–1161.

Aumann, R.J., and M. Maschler. 1985. Game theoretic analysis of a bankruptcy problem from the Talmud. *Journal of Economic Theory* 36: 195–213.

Aumann, R.J., and B. Peleg. 1960. Von Neumann-Morgenstern solutions to cooperative games without side payments. *Bulletin of the American Mathematical Society* 66: 173–179.

Aumann, R.J., and L.S. Shapley. 1974. *Values of non-atomic games*. Princeton: Princeton University Press.

Axelrod, R. 1984. *The evolution of cooperation*. New York: Basic Books.

Balinski, M.L., and H.P. Young. 1982. *Fair representation*. New Haven: Yale University Press.

Berbee, H. 1981. On covering single points by randomly ordered intervals. *Annals of Probability* 9: 520–528.

Bewley, T., and E. Kohlberg. 1976. The asymptotic theory of stochasticgames. *Mathematics of Operations Research* 1: 197–208.

Billera, L.J. 1970a. Existence of general bargaining sets for cooperative games without side payments. *Bulletin of the American Mathematical Society* 76: 375–379.

Billera, L.J. 1970b. Some theorems on the core of an n-person game without side payments. *SIAM Journal of Applied Mathematics* 18: 567–579.

Billera, L.J., and R. Bixby. 1973. A characterization of polyhedral market games. *International Journal of Games Theory* 2: 253–261.

Billera, L.J., and D.C. Heath. 1982. Allocation of shared costs: A set of axioms yielding a unique procedure. *Mathematics of Operations Research* 7: 32–39.

Billera, L.J., D.C. Heath, and J. Raanan. 1978. Internal telephone billing rates – A novel application of non-atomic game theory. *Operations Research* 26: 956–965.

Binmore, K. 1982. *Perfect equilibria in bargaining models*, ICERD discussion paper no. 58. London: London School of Economics.

Blackwell, D. 1956. An analogue of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics* 6: 1–8.

Blackwell, D., and T.S. Ferguson. 1968. The big match. *Annals of Mathematical Statistics* 39: 159–163.

Bogardi, I., and F. Szidarovsky. 1976. Application of game theory in water management. *Applied Mathematical Modelling* 1: 11–20.

Bondareva, O.N. 1963. Some applications of linear programming methods to the theory of cooperative games [in Russian]. *Problemy Kibernetiki* 10: 119–139.

Borel, E. 1924. Sur les jeux où interviennent l'hasard et l'habilité des joueurs. In *Eléments de la theorie des probabilités*, ed. J. Hermann, 204–224. Paris: Librairie Scientifique.

Braithwaite, R.B., and F.P. Ramsey, eds. 1950. *The foundations of mathematics and other logical essays*. New York: Humanities Press.

Brams, S.J., W.F. Lucas, and P.D. Straffin Jr., eds. 1983. *Political and related models*. New York: Springer.

Brown, D.J., and P. Loeb. 1976. The values of nonstandard exchangeeconomies. *Israel Journal of Mathematics* 25: 71–86.

Brown, D.J., and A. Robinson. 1975. Non standard exchange economies. *Econometrica* 43: 41–55.

Case, J.H. 1979. *Economics and the competitive process*. New York: New York University Press.

Champsaur, P. 1975. Cooperation vs. competition. *Journal of Economic Theory* 11: 394–417.

Dantzig, G.B. 1951a. A proof of the equivalence of the programming problem and the game problem. In *Koopmans* (1951), 330–338.

Dantzig, G.B. 1951b. Maximization of a linear function of variables subject to linear inequalities. In *Koopmans* (1951), 339–347.

Davis, M. 1964. Infinite games with perfect information. In *Dresher, Shapley and Tucker* (1964), 85–101.

Davis, M. 1967. Existence of stable payoff configurations for cooperative games. In *Shubik* (1967), 39–62.

Davis, M., and M. Maschler. 1965. The kernel of a cooperative game. *Naval Research Logistics Quarterly* 12: 223–259.

Debreu, G. 1952. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences of the United States* 38: 886–893.

Debreu, G., and H. Scarf. 1963. A limit theorem on the core of aneconomy. *International Economic Review* 4: 236–246.

Dresher, M.A., A.W. Tucker, and P. Wolfe, eds. 1957. *Contributions to the theory of games III*, Annals of Mathematics studies series 39. Princeton: Princeton University Press.

Dresher, M.A., L.S. Shapley, and A.W. Tucker, eds. 1964. *Advances in game theory*, Annals of Mathematics studies series 52. Princeton: Princeton University Press.

Drèze, J.H., Gabszewicz, J., and Gepts, S. 1969. On cores and competitive equilibria. In *Guilbaud* (1969), 91–114.

Dubey, P. 1980. Asymptotic semivalues and a short proof of Kannai's theorem. *Mathematics of Operations Research* 5: 267–270.

Dubey, P., and Shapley, L.S. 1980. Non cooperative exchange with a continuum of traders: Two models. *Technical Report of the Institute for Advanced Studies*, Hebrew University of Jerusalem.

Edgeworth, F.Y. 1881. *Mathematical psychics*. London: Kegan Paul.

Everett, H. 1957. Recursive games. In *Dresher, Tucker and Wolfe* (1957), 47–78.

Fogelman, F., and M. Quinzii. 1980. Asymptotic values of mixedgames. *Mathematics of Operations Research* 5: 86–93.

Gabszewicz, J.J., and J.F. Mertens. 1971. An equivalence theorem for the core of an economy whose atoms are not 'too' big. *Econometrica* 39: 713–721.

Gale, D. 1974. A curious nim-type game. *American Mathematical Monthly* 81: 876–879.

Gale, D. 1979. The game of hex and the Brouwer fixed-point theorem. *American Mathematical Monthly* 86: 818–827.

Gale, D. 1986. Bargaining and competition, Part I: Characterization. Part II: Existence. *Econometrica* 54 (785–806): 807–818.

Gale, D., and L.S. Shapley. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69: 9–15.

Gale, D., and F.H. Stewart. 1953. Infinite games with perfect information. In *Kuhn and Tucker* (1953), 245–266.

Galil, Z. 1974. The nucleolus in games with major and minor players. *International Journal of Games Theory* 3: 129–140.

Gillies, D.B. 1959. Solutions to general non-zero-sum games. In *Luce and Tucker* (1959), 47–85.

Guilbaud, G.T., ed. 1969. *La décision: aggrégation et dynamique des orders de préférence*. Paris: Editions du CNRS.

Harsanyi, J.C. 1956. Approaches to the bargaining problem before and after the theory of games: A critical discussion of Zeuthen's, Hicks' and Nash's theories. *Econometrica* 24: 144–157.

Harsanyi, J.C. 1959. A bargaining model for the cooperative n-person game. In *Tucker and Luce* (1959), 325–356.

Harsanyi, J.C. 1963. A simplified bargaining model for the n-person cooperative game. *International Economic Review* 4: 194–220.

Harsanyi, J.C. 1966. A general theory of rational behavior in game situations. *Econometrica* 34: 613–634.

Harsanyi, J.C. 1967–8. Games with incomplete information played by'Bayesian' players, parts I, II and III. *Management Science* 14: 159–182, 320–334, 486–502.

Harsanyi, J.C. 1973. Games with randomly disturbed payoffs: A new rationale for mixed strategy equilibrium points. *International Journal of Games Theory* 2: 1–23.

Harsanyi, J.C. 1982. Solutions for some bargaining games under the Harsanyi-Selten solution theory I: Theoretical preliminaries; II: Analysis of specific games. *Mathematical Social Sciences* 3 (179–91): 259–279.

Harsanyi, J.C., and R. Selten. 1972. A generalized Nash solution for two-person bargaining games with incomplete information. *Management Science* 18: 80–106.

Harsanyi, J.C., and R. Selten. 1987. *A general theory of equilibrium selection in games*. Cambridge, MA: MIT Press.

Hart, S. 1973. Values of mixed games. *International Journal of Games Theory* 2: 69–86.

Hart, S. 1974. Formation of cartels in large markets. *Journal of Economic Theory* 7: 453–466.

Hart, S. 1977a. Asymptotic values of games with a continuum of players. *Journal of Mathematical Economics* 4: 57–80.

Hart, S. 1977b. Values of non-differentiable markets with a continuum of traders. *Journal of Mathematical Economics* 4: 103–116.

Hart, S. 1980. Measure-based values of market games. *Mathematics of Operations Research* 5: 197–228.

Hart, S. 1985a. An axiomatization of Harsanyi's non-transferable utility solution. *Econometrica* 53: 1295–1314.

Hart, S. 1985b. Non zero-sum two-person repeated games with incomplete information. *Mathematics of Operations Research* 10: 117–153.

Hart, S., and A. Mas-Colell. 1986. The potential: A new approach to the value in multi-person allocation problems. Harvard University Discussion Paper No. 1157.

Hart, S., and D. Schmeidler. 1988. Correlated equilibria: An elementary existence proof. *Mathematics of Operations Research*.

Hildenbrand, W., ed. 1982. *Advances in economic theory*. Cambridge: Cambridge University Press.
Hildenbrand, W., and A. Mas-Colell. 1986. *Contributions to mathematical economics in honor of G. Debreu*. Amsterdam: North-Holland.
Hu, T.C., and S.M. Robinson, eds. 1973. *Mathematical programming*. New York: Academic.
Isaacs, R. 1965. *Differential games: A mathematical theory with applications to warfare and pursuit, control and optimization*. New York: Wiley.
Kakutani, S. 1941. A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal* 8: 457–459.
Kalai, E., and M. Smorodinsky. 1975. Other solutions to Nash's bargainingproblem. *Econometrica* 43: 513–518.
Kaneko, M., and M. Wooders. 1982. Cores of partitioning games. *Mathematical Social Sciences* 3: 313–327.
Kannai, Y. 1966. Values of games with a continuum of players. *Israel Journal of Mathematics* 4: 54–58.
Kohlberg, E. 1971. On the nucleolus of a characteristic function game. *SIAM Journal of Applied Mathematics* 20: 62–66.
Kohlberg, E. 1972. The nucleolus as a solution to a minimization problem. *SIAM Journal of Applied Mathematics* 23: 34–49.
Koopmans, T.C., ed. 1951. *Activity analysis of production and allocation*. New York: Wiley.
Kuhn, H.W. 1952. *Lectures on the theory of games*. Issued as a report of the Logistics Research Project, Office of Naval Research, Princeton University.
Kuhn, H.W. 1953. Extensive games and the problem of information. In *Kuhn and Tucker* (1953), 193–216.
Kuhn, H.W., and A.W. Tucker, eds. 1950. *Contributions to the theory of games I*, Annals of Mathematics studies series 24. Princeton: Princeton University Press.
Kuhn, H.W., and A.W. Tucker, eds. 1953. *Contributions to the theory of games II*, Annals of Mathematics studies series 28. Princeton: Princeton University Press.
Lemke, L.E., and J.T. Howson. 1962. Equilibrium points of bimatrix games. *SIAM Journal of Applied Mathematics* 12: 413–423.
Lensberg, T. 1981. The stability of the Nash solution. Unpublished.
Lewis, D.K. 1969. *Convention*. Cambridge, MA: Harvard University Press.
Littlechild, S.C. 1976. A further note on the nucleolus of the 'airportgame'. *International Journal of Games Theory 5*: 91–95.
Littlechild, S.C., and G. Owen. 1973. A simple expression for the Shapley value in a special case. *Management Science* 20: 370–372.
Lucas, W.F. 1969. The proof that a game may not have a solution. *Transactions of the American Mathematical Society* 137: 219–229.
Lucas, W.F. 1983. Measuring power in weighted voting systems. In *Brams, Lucas and Straffin* (1983), ch. 9.
Lucas, W.F., and M. Rabie. 1982. Games with no solutions and emptycore. *Mathematics of Operations Research* 7: 491–500.
Luce, R.D., and H. Raiffa. 1957. *Games and decisions, introduction and critical survey*. New York: Wiley.
Luce, R.D., and A.W. Tucker, eds. 1959. *Contributions to the theory of games IV*, Annals of Mathematics studies series 40. Princeton: Princeton University Press.
Lyapounov, A.A. 1940. On completely additive vector-functions (in Russian, abstract in French). *Akademiia Nauk USSR Izvestiia Seriia Mathematicheskaia* 4: 465–478.
Martin, D.A. 1975. Borel determinacy. *Annals of Mathematics* 102: 363–371.
Maschler, M., ed. 1962. *Recent advances in game theory*. Proceedings of a Conference, privately printed for members of the conference. Princeton: Princeton University Conferences.
Maschler, M., and M. Perles. 1981. The superadditive solution for the Nash bargaining game. *International Journal of Games Theory* 10: 163–193.
Maschler, M., B. Peleg, and L.S. Shapley. 1979. Geometric properties of the kernel, nucleolus, and related solution concepts. *Mathematics of Operations Research* 4: 303–338.
Mas-Colell, A. 1975. A further result on the representation of games by markets. *Journal of Economic Theory* 10: 117–122.
Mas-Colell, A. 1977. Competitive and value allocations of large exchange economies. *Journal of Economic Theory* 14: 419–438.
Mas-Colell, A. 1988. An equivalence theorem for a bargaining set. *Journal of Mathematical Economics*.
Maynard Smith, J. 1982. *Evolution and the theory of games*. Cambridge: Cambridge University Press.
Mertens, J.F. 1982. Repeated games: An overview of the zero-sum case. In *Hildenbrand* (1982), 175–182.
Mertens, J.F. 1988. The Shapley value in the non-differentiable case. *International Journal of Games Theory* 17: 1–65.
Mertens, J.F., and A. Neyman. 1981. Stochastic games. *International Journal of Games Theory* 10: 53–66.
Milnor, J.W. 1978. Values of large games II: Oceanic games. *Mathematics of Operations Research* 3: 290–307.
Milnor, J.W. and Shapley, L.S. 1957. On games of survival. In *Dresher, Tucker and Wolfe* (1957), 15–45.
Moschovakis, Y.N. 1980. *Descriptive set theory*. New York: North-Holland.
Moschovakis, Y.N., ed. 1983. *Cabal seminar 79–81: Proceedings, Caltech-UCLA Logic Seminar 1979–81*, Lecture notes in Mathematics 1019. New York: Springer-Verlag.
Mycielski, J., and H. Steinhaus. 1964. On the axiom of determinateness. *Fundamenta Mathematicae* 53: 205–224.
Myerson, R.B. 1977. Graphs and cooperation in games. *Mathematics of Operations Research* 2: 225–229.

Myerson, R.B. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47: 61–74.

Myerson, R.B. 1984. Cooperative games with incomplete information. *International Journal of Games Theory* 13: 69–96.

Nash, J.F. Jr. 1950. The bargaining problem. *Econometrica* 18: 155–162.

Nash, J.F. Jr. 1951. Non-cooperative games. *Annals of Mathematics* 54: 289–295.

Neyman, A. 1985a. Bounded complexity justifies cooperation in the finitely repeated prisoner's dilemma. *Economics Letters* 19: 227–230.

Neyman, A. 1985b. Semivalues of political economic games. *Mathematics of Operations Research* 10: 390–402.

Neyman, A. 1987. Weighted majority games have an asymptotic value. *Mathematics of Operations Research*.

O'Neill, B. 1987. Non-metric test of the minimax theory of two-person zero-sum games. *Proceedings of the National Academy of Sciences of the United States* 84: 2106–2109.

Owen, G. 1972. Multilinear extensions of games. *Management Science* 18: 64–79.

Peleg, B. 1963a. Solutions to cooperative games without side payments. *Transactions of the American Mathematical Society* 106: 280–292.

Peleg, B. 1963b. Bargaining sets of cooperative games without side payments. *Israel Journal of Mathematics* 1: 197–200.

Peleg, B. 1967. Existence theorem for the bargaining set M1(i). In *Shubik* (1967), 53–56.

Peleg, B. 1968. On weights of constant-sum majority games. *SIAM Journal of Applied Mathematics* 16: 527–532.

Peleg, B. 1985. An axiomatization of the core of cooperative games without side payments. *Journal of Mathematical Economics* 14: 203–214.

Peleg, B. 1986. On the reduced game property and its converse. *International Journal of Games Theory* 15: 187–200.

Pennock, J.R., and J.W. Chapman, eds. 1968. *Representation*. New York: Atherton.

Ramsey, F.P. 1931. Truth and probability. In Braithwaite (1950).

Ransmeier, J.S. 1942. *The Tennesee valley authority: A case study in the economics of multiple purpose stream planning*. Nashville: Vanderbilt University Press.

Regelmann, K., and E. Curio. 1986. How do great tit (Parus major) pair mates cooperate in broad defence? *Behavior* 97: 10–36.

Riker, W.H. and Shapley, L.S. 1968. Weighted voting: A mathematical analysis for instrumental judgements. In *Pennock and Chapman* (1968), 199–216.

Robinson, A. 1974. *Non-standard analysis*. Amsterdam: North-Holland.

Rosenthal, R.W. 1979. Sequences of games with varying opponents. *Econometrica* 47: 1353–1366.

Rosenthal, R.W., and A. Rubinstein. 1984. Repeated two player games withruin. *International Journal of Games Theory* 13: 155–177.

Roth, A.E. 1977. The Shapley value as a von Neumann-Morgenstern utility. *Econometrica* 45: 657–664.

Roth, A.E. 1984. The evolution of the labor market for medical interns and residents: A case study in game theory. *Journal of Political Economy* 92: 991–1016.

Roth, A.E., and R.E. Verrecchia. 1979. The Shapley value as applied to cost allocation: A reinterpretation. *Journal of Accounting Research* 17: 295–303.

Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50: 97–109.

Rubinstein, A. 1986. Finite automata play the repeated prisoner's dilemma. *Journal of Economic Theory* 39: 83–96.

Savage, L.J. 1954. *The foundations of statistics*. New York: Wiley.

Scarf, H.E. 1967. The core of an n-person game. *Econometrica* 35: 50–69.

Scarf, H.E. 1973. *The computation of economic equilibria*. New Haven: Yale University Press.

Schelling, T.C. 1960. *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Schmeidler, D. 1969. The nucleolus of a characteristic function game. *SIAM Journal of Applied Mathematics* 17: 1163–1170.

Selten, R.C. 1965. Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit. *Zeitschrift für die Gesammte Staatswissenschaft* 121: 301–324; 667–689.

Selten, R.C. 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Games Theory* 4: 25–55.

Selten, R.C. 1980. A note on evolutionary stable strategies in asymmetric animal conflicts. *Journal of Theoretical Biology* 84: 101–101.

Selten, R.C. 1983. Evolutionary stability in extensive two-part games. *Mathematical Social Sciences* 5: 269–363.

Shapley, L.S. 1953a. Stochastic games. *Proceedings of the National Academy of Sciences of the United States* 39: 1095–1100.

Shapley, L.S. 1953b. A value for n-person games. In *Kuhn and Tucker* (1953), 305–317.

Shapley, L.S. 1962. Values of games with infinitely many players. In *Maschler* (1962), 113–118.

Shapley, L.S. 1964. Values of large games, VII: A general exchange economy with money. *RAND Publication RM–4248*, Santa Monica.

Shapley, L.S. 1967. On balanced sets and cores. *Naval Research Logistics Quarterly* 14: 453–460.

Shapley, L.S. 1969a. Utility comparison and the theory of games. In *Guilbaud* (1969), 251–263.

Shapley, L.S. 1969b. On market games. *Journal of Economic Theory* 1: 9–25.

Shapley, L.S. 1969c. Pure competition, coalitional power and fair division. *International Economic Review* 10: 337–362.

Shapley, L.S. 1973a. On balanced games without side payments. In *Hu and Robinson* (1973), 261–290.
Shapley, L.S. 1973b. Let's block 'block'. *Econometrica* 41: 1201–1202.
Shapley, L.S. 1984. Convergence of the bargaining set for differentiable market games. Appendix B in Shubik (1984), 683–692.
Shapley, L.S., and M. Shubik. 1954. A method for evaluating the distribution of power in a committee system. *American Political Science Review* 48: 787–792.
Shitovitz, B. 1973. Oligopoly in markets with a continuum of traders. *Econometrica* 41: 467–501.
Shubik, M. 1959a. Edgeworth market games. In *Luce and Tucker* (1959), 267–278.
Shubik, M. 1959b. *Strategy and market structure*. New York: Wiley.
Shubik, M., ed. 1967. *Essays in mathematical economics in honor of Oskar Morgenstern*. Princeton: Princeton University Press.
Shubik, M. 1973. Commodity, money, oligopoly, credit and bankruptcy in a general equilibrium model. *Western Economic Journal* 11: 24–36.
Shubik, M. 1982. *Game theory in the social sciences, concepts and solutions*. Cambridge, MA: MIT Press.
Shubik, M. 1984. *A game theoretic approach to political economy*. Cambridge, MA: MIT Press.
Sobolev, A.I. 1975. Characterization of the principle of optimality for cooperative games through functional equations (in Russian). In *Vorobiev* (1975), 94–151.
Sorin, S. 1986a. On repeated games of complete information. *Mathematics of Operations Research* 11: 147–160.
Sorin, S. 1986b. An asymptotic property of non-zero sum stochastic games. *International Journal of Games Theory* 15 (2): 101–107.
Tauman, Y. 1981. Value on a class of non-differentiable market games. *International Journal of Games Theory* 10: 155–162.
Ville, J.A. 1938. Sur le théorie générale des jeux où intervient l'habilité des joueurs. In *Traité du calcul des probabilités et de ses applications*, ed. E. Borel, vol. 4, 105–113. Paris: Gauthier-Villars.
Vinacke, W.E., and A. Arkoff. 1957. An experimental study of coalitions in the triad. *American Sociological Review* 22: 406–415.
Vind, K. 1965. A theorem on the core of an economy. *Review of Economic Studies* 32: 47–48.
von Neumann, J. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100: 295–320.
von Neumann, J. 1949. On rings of operators. Reduction theory. *Annals of Mathematics* 50: 401–485.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.
Vorobiev, N.N., ed. 1975. *Mathematical Methods in Social Science* (in Russian). Vipusk 6, Vilnius.
Wilson, R. 1978. Information, efficiency, and the core of an economy. *Econometrica* 46: 807–816.
Wolfe, P. 1955. The strict determinateness of certain infinite games. *Pacific Journal of Mathematics* 5: 841–847.
Wooders, M.H. 1983. The epsilon core of a large replica game. *Journal of Mathematical Economics* 11: 277–300.
Wooders, M.H., and W.R. Zame. 1984. Approximate cores of largegames. *Econometrica* 52: 1327–1350.
Young, H.P. 1985. Monotonic solutions of cooperative games. *International Journal of Games Theory* 14: 65–72.
Zermelo, E. 1913. Über eine Anwendung der Mengenlehre auf die theorie des Schachspiels. *Proceedings of the Fifth International Congress of Mathematicians* 2: 501–504.

Game Theory and Biology

Olof Leimar

Abstract

Darwinian evolutionary dynamics and learning dynamics provide the foundation for game theory in biology. The theory is used to analyse interactions between individuals. Animal fighting behaviour, cooperative interactions and signalling interactions are examples of important areas of application. The payoffs to strategies in biological games represent Darwinian fitness, viz. survival and reproductive success. The strategies can be behaviour patterns, but also choices of phenotypic properties such as becoming a male or a female. The evolutionary analysis of allocation to male and female function is one of the most successful applications of game theory in biology.

Keywords

Alternative reproductive strategies; Class structured populations; Cooperation; Deterministic evolutionary dynamics; Division of labour; Evolutionarily stable strategies; Evolutionary theory; Game theory and biology; Learning and evolution in games; Life-history strategy; Mate choice; Mixed strategy equilibria; Natural selection; Playing the field;

Prisoner's Dilemma; Producer–scrounger game; Reciprocity; Reproductive value; Sex ratio theory; Signalling; Tit for tat

JEL Classifications
C7

In biology, game theory is a branch of evolutionary theory that is particularly suited to the study of interactions between individuals. The evolution of animal fighting behaviour was among the first applications and it was in this context that Maynard Smith and Price (1973) developed the concept of an evolutionarily stable strategy (ESS) (see ▶ Learning and Evolution in Games: ESS). Cooperative interactions (Trivers 1971) and signalling interactions (Grafen 1991), such as when males signal their quality to females, are examples of other important areas of application. There is an overlap of ideas between economics and biology, which has been quite noticeable since the 1970s and, in a few instances, earlier (Sigmund 2005). In the early 21st century, the interchange takes the form of a joint exploration of theoretical and empirical issues by biologists and economists (Hammerstein and Hagen 2005).

Strategies in games inspired by biology can represent particular behaviour patterns, including rules about which behaviour to perform in which circumstance. Other aspects of an individual's phenotype can also be viewed as the result of strategic choice. A life-history strategy specifies choices that have major impact on an individual's course of life, for instance, whether to become a male or a female or, for certain insects, whether or not to develop wings. Interactions between individuals are modelled as games where the payoffs represent Darwinian fitness. Random matching of players drawn from a large population is one common game model, which was used to study fighting behaviour (Maynard Smith and Price 1973; Maynard Smith and Parker 1976). 'Playing the field' (Maynard Smith 1982) is a more general modelling approach, where the payoff to an individual adopting a particular strategy depends on some average property of the population (cf. population games in deterministic evolutionary dynamics).

Game theory is needed for situations where payoffs to strategies depend on the state of a population, and this state in turn depends on the strategies that are present. For matching of players drawn from a population, the distribution of opposing strategies is of course given by the population distribution, but there are other reasons why the distribution of strategies influences expected payoffs. A 'playing-the-field' example is the choice by an individual, or by its mother, to develop into a male or a female. The two sexes occur in roughly equal proportions in many species. This observation intrigued Darwin, who was unable to provide a satisfactory explanation, writing that 'I formerly thought that when a tendency to produce the two sexes in equal numbers was advantageous to the species, it would follow from natural selection, but I now see that the whole problem is so intricate that it is safer to leave its solution to the future' (Darwin 1874, p. 399). The solution to the problem was found by Düsing (1884; see also Edwards 2000, and Fisher 1930), and rested on the principle that, in diploid sexual organisms, the total genetic contribution to offspring by all males in a generation must equal the contribution by all females in the same generation. This gives a reproductive advantage to the rarer sex in the passing of genes to future generations. The payoffs to a mother from producing a son or a daughter must then depend on the population sex ratio, and this dependence can result in an evolutionary equilibrium at an even sex ratio (see below). The idea arose before the development of the concept of an ESS by Maynard Smith and Price (1973), but it can be regarded as the first instance of game-theoretical reasoning in biology.

Payoffs, Reproductive Value, and Evolutionary Dynamics

Class-structured populations in discrete time (Caswell 2001) are often used as settings for evolutionary analysis. The classes or states are properties like female and male, and time might be measured in years. Let $n_i(t)$ denote the number of individuals in state i at time t. We can write a deterministic population dynamics as $n(t + 1)$

$=An(t)$, where n is the vector of the n_i and A is the so-called population projection matrix. The elements a_{ij} of A can depend on n and on the strategies that are present in the population. They represent per capita genetic contributions of individuals in state j to state i, in terms of offspring or individual survival. A common evolutionary analysis is to determine a stationary n for the case where all individuals use a strategy x and to examine whether rare mutants with strategy x' would increase in number in such a population.

Let us apply this scheme to the just mentioned sex ratio problem. Suppose that a mother can determine the sex of her offspring and that females in the population produce a son with probability x and a daughter with probability $1 - x$. For nonoverlapping generations, the dynamics $n(t + 1) = An(t)$ can be written as

$$\begin{pmatrix} n_f(t+1) \\ n_m(t+1) \end{pmatrix} = \begin{pmatrix} 0.5(1-x)b & 0.5(1-x)bq \\ 0.5xb & 0.5xb \end{pmatrix} \begin{pmatrix} n_f(t) \\ n_m(t) \end{pmatrix},$$

where b is the reproductive output (number of offspring) of a female, bq is the reproductive output of a male, and the factor 0.5 accounts for the genetic shares of the parents in their offspring. Because the reproductive output of all males must equal that of all females, it follows that $q = n_f(t)/n_m(t)$ and thus that $q = (1 - x)/x$. In a stationary population, $b = 1/(1 - x)$ must hold, which could come about through a dependence of b on the total population size. Introducing the matrix

$$B(x', x) = \frac{1}{2} \begin{pmatrix} (1-x')/(1-x) & (1-x)/x \\ x'/(1-x) & 1 \end{pmatrix},$$

the population projection matrix for a stationary population is $A = B(x, x)$ and the stationary $n = (n_f, n_m)$ is proportional to the leading eigenvector, $w = (1 - x, x)$, of this A. Suppose a mutant gene causes a female to adopt the sex ratio strategy x', but has no effect in a male. As long as the mutant gene is rare, only the strategy of heterozygous mutant females needs to be taken into account, and the dynamics of the mutant sub-population can be written as $n'(t + 1) = A'n'(t)$ with $A' = B(x', x)$. The mutant can invade if $\lambda(x', x) > 1$ holds for the leading eigenvalue $\lambda(x', x)$ of $B(x', x)$. Direct computation of the leading eigenvalue shows that a mutant with $x' > x$ can invade if $x < 0.5$ and one with $x' < x$ can invade if $x' > 0.5$, resulting in an evolutionary equilibrium at $x = 0.5$.

The reproductive value of state i is defined as the ith component of the leading 'left eigenvector' v of the stationary population projection matrix $A = B(x, x)$, that is, v is the leading eigenvector of the transpose of A. It is convenient to normalize v so that its scalar product $v \cdot w$ with the leading eigenvector w equals 1. For our sex ratio problem we have $v = \frac{1}{2}(1/(1 - x), 1/x)$. The reproductive value of state i can be interpreted as being proportional to the expected genetic contribution to future generations of an individual in state i. The eigenvectors v and w can be used to investigate how the leading eigenvalue depends on x' near $x' = x$. It is easy to show that $\partial\lambda(x', x)/\partial x' = \partial(v \cdot B(x', x)w)/\partial x'$ holds at $x' = x$ (for example, Caswell 2001), and this result can be used to identify evolutionary equilibria. If a mutation has an effect in only one of the states, like the females in our example, there is further simplification in that only one column of $B(x', x)$ depends on x'. It follows that evolutionary change through small mutational steps in the sex ratio example can be described as if females were selected to maximize the expected reproductive value per offspring, given by $V(x', x) = \frac{1}{2}(1 - x')/(1 - x) + \frac{1}{2}x'/x$. Payoff functions having this form were introduced by Shaw and Mohler (1953), in what may have been the first worked out game-theoretical argument in biology. As we have seen, analysis of such payoff functions corresponds to an analysis of mutant invasion in a stationary population.

The concept of reproductive value was introduced by Fisher (1930) and plays an important role in the very successful field of sex ratio theory (Charnov 1982; Pen and Weissing 2002), as well as in evolutionary theory in general (McNamara and Houston 1996; Houston and McNamara 1999; Grafen 2006). The concept is useful to represent payoffs in games played in populations in stationary environments. Reproductive value can be regarded as a Darwinian representation of the concept of utility in economics. For populations exposed to

large-scale environmental fluctuations, as well as for those with limit-cycle or chaotic attractors of the population dynamics, concepts similar to reproductive value have proven less useful. In such situations, one needs the more general approach of explicitly considering evolutionary dynamics for populations of players of strategies. There are several influential approaches to the study of evolutionary dynamics in biology (Nowak and Sigmund 2004), ranging from replicator dynamics (see deterministic evolutionary dynamics) and adaptive dynamics (Metz et al. 1992; Metz et al. 1996; Hofbauer and Sigmund 1998) to the traditional modelling styles of population genetics and quantitative genetics (Rice 2004). These approaches make different assumptions about such things as the underlying genetics and the rate and distribution of mutation. Recent years have seen an increasing emphasis on explicitly dynamical treatments in evolutionary theory.

Are There Mixed Strategies in Nature?

Biologists have wondered how individuals, as players of a game, come to play one strategy or another. For life-history strategies, involving choices between alternative phenotypes, a population containing a mixture of phenotypes could be the result of randomization at the level of an individual, which corresponds to a mixed strategy, or there could be a genetic polymorphism of pure strategies (Maynard Smith 1982). These two possibilities can be contrasted with a third, where individuals (or their parents) use information about themselves or their local environment to make life-history choices, which could correspond to a conditional strategy in a Bayesian game. The general question is related to the issue of purification of mixed strategy equilibria in game theory (see ▶ Purification). When observing populations that are mixtures of discrete phenotypes, biologists have tried to establish if one of the above three possibilities applies and, if so, what the evolutionary explanation might be. This question has been asked, for instance, about the phenomenon of alternative reproductive strategies (Gross 1996; Shuster and Wade 2003), like the jack and hooknose males in coho salmon (Gross 1985) or the winged and wingless males in fig wasps (Hamilton 1979). Since there are likely to be a number of factors that influence the relative success of reproductive alternatives and could be known to a developing individual – for instance, its juvenile growth rate and thus its potential adult size – one might expect some form of conditional strategy to evolve. This expectation agrees with the observation that conditional determination is common (Gross 1996). There are also instances of genetic determination of reproductive alternatives (Shuster and Wade 2003) but, somewhat surprisingly, there is as yet no empirically confirmed case of a mixed strategy of this kind. Perhaps it has been difficult for evolution to construct a well-functioning randomization device, leaving genetic polymorphism as a more easily achieved evolutionary outcome.

Evolution of Cooperation

Among the various applications of game theory in biology, the evolution of cooperation is by far the most studied issue. This great interest is based on the belief that cooperation has played a crucial role in the evolution of biological organization, from the structure of chromosomes, cells and organisms to the level of animal societies. An extreme form of cooperation is that of the genes operating in an organism. Several thousand genes coordinate and direct cellular activities that in the main serve the well-being of their organism. Kin selection (Hamilton 1964), which predicts that agents have an evolutionary interest in assisting their genetic relatives, cannot be the main explanation for this cooperation, since the different genes in an organism are typically not closely related by descent (except for a given gene in one cell and its copies in other cells). It is instead division of labour that is the principle that unites the parts of an organism into a common interest, of sufficient strength to make it evolutionarily unprofitable for any one gene to abandon its role in the organism for its own advantage. There are of course exceptions, in the form of selfish genetic

elements, but these represent a minority of cases (Burt and Trivers 2006).

Trivers (1971) and Axelrod and Hamilton (1981) promoted the idea that many of the features of the interactions between organisms would find an explanation in the give and take of direct reciprocation. In particular, the strategy of tit for tat (Axelrod and Hamilton 1981) for the repeated Prisoner's Dilemma game was thought to represent a general mechanism for reciprocity in cooperative interactions and received much attention from biologists. On the whole, this form of direct reciprocity has subsequently failed to be supported by empirical observation. Two reasons for this failure have been proposed (Hammerstein 2003). One is that the structure of real biological interactions differs in important ways from the original theoretical assumptions of a repeated game. The other is that the proposed strategies, like tit for tat, are unlikely to be reached by evolutionary change in real organisms, because they correspond to unlikely behavioural mechanisms. In contrast to reciprocity, both the influence of genetic relatedness through kin selection (Hamilton 1964) and the presence of direct fitness benefits to cooperating individuals have relatively strong empirical support. Division of labour and the direct advantages of the trading of benefits between agents are likely to be crucial ingredients in the explanation of cooperation between independent organisms. The idea of a market, where exchanges take place, is thus relevant in both biology and economics (Noe and Hammerstein 1994).

Evolution of Signalling

Signals are found in a wide variety of biological contexts, for instance in aggressive interactions, parent–offspring interactions, and in connection with mate choice. There is now a fairly well developed set of theories about biological signals (Maynard Smith and Harper 2003). One of the most influential ideas in the field is Zahavi's handicap principle (Zahavi 1975). It states that a signal can reliably indicate high quality of the signaller only if the signal is costly, to the extent that it does not pay low-quality individuals to display the signal. The idea can be seen as a nonmathematical version of Spence's signalling theory (Spence 1973, 1974), but, because biologists, including Zahavi, were unaware of Spence's work in economics, Zahavi's principle remained controversial in biology until Grafen (1991) provided a game-theoretical justification. The turn of events illustrates that biologists might have benefited from being more aware of theoretical developments in economics.

An example where Zahavi's handicap principle could apply is female mate choice in stalk-eyed flies (David et al. 2000). Males of stalk-eyed flies have long eye stalks, increasing the distance between their eyes, which is likely to be an encumbrance in their day-to-day life. A high level of nutrition, but also the possession of genes for high phenotypic quality, cause males to develop longer eye stalks. Female stalk-eyed flies prefer to mate with males with eyes that are far apart, and in this way their male offspring have a greater chance of receiving genes for long eye stalks. Female choice will act to reduce genetic variation in males, but if a sufficiently broad range of genetic loci can influence eye-stalk length, because they have a general effect on the phenotypic quality of a male, processes like deleterious mutation could maintain a substantial amount of genetic variation. In this way, signalling theory can explain the evolution of elaborate male ornaments, together with a mating preference for these ornaments in females, illustrating the power of game-theoretical arguments to increase our understanding of biological phenomena.

Learning

Viewing strategies as genetically coded entities on which natural selection operates, with evolutionarily stable strategies as endpoints of evolutionary change, is not the only game-theoretical perspective that is of relevance in biology. For many categories of behaviour, learning or similar adjustment processes are important in shaping the distribution of strategies in a population. For instance, when animals search for food or locate suitable living quarters, they may have the

opportunity to evaluate the relative success of different options and to adjust their behaviour accordingly. A well-studied example is the so-called producer–scrounger game, for which there are experiments with birds that forage in groups on the ground (Barnard and Sibly 1981; Giraldeau and Caraco 2000). The game is played by a group of foragers and consists of a number of rounds. In each round an individual can choose between two behavioural options. Producers search for and utilize new food sources, and scroungers exploit food found by producers. The game presupposes that the activities of producing and scrounging are incompatible and cannot be performed simultaneously, which is experimentally supported (Coolen et al. 2001). The payoffs to the options, measured as the expected food intake per round, depend on the frequencies of the options in the group. For instance, scrounging is most profitable to an individual if no one else scrounges and yields a lower payoff with more scrounging in the group. By specifying the details of the model, one can compute an equilibrium probability of scrounging at which the payoffs to producing and scrounging are equal. This equilibrium is influenced by parameters like expected search times and the amount of food found in a new location. It has been experimentally verified that groups of spice finches converge on such an equilibrium over a period of a few days of foraging (Mottley and Giraldeau 2000; Giraldeau and Caraco 2000). It is not known precisely which rules are used by individuals in these experiments to modify their behaviour, but such rules are likely to play an important role in shaping behaviour in many animals, including humans (see ▶ Learning and Evolution in Games: Adaptive Heuristics). The study of these kinds of adjustments of behaviour could therefore represent an important area of overlap between biology and economics.

See Also

- ▶ Deterministic Evolutionary Dynamics
- ▶ Learning and Evolution in Games: Adaptive Heuristics
- ▶ Learning and Evolution in Games: ESS
- ▶ Mixed Strategy Equilibrium
- ▶ Purification
- ▶ Utility

Bibliography

Axelrod, R., and W.D. Hamilton. 1981. The evolution of cooperation. *Science* 211: 1390–1396.

Barnard, C.J., and R.M. Sibly. 1981. Producers and scroungers: A general model and its application to captive flocks of house sparrows. *Animal Behaviour* 29: 543–550.

Burt, A., and R. Trivers. 2006. *Genes in conflict: the biology of selfish genetic elements*. Cambridge: Harvard University Press.

Caswell, H. 2001. *Matrix population models*. 2nd ed. Sunderland: Sinauer.

Charnov, E.L. 1982. *The theory of sex allocation*. Princeton: Princeton University Press.

Coolen, I., L.-A. Giraldeau, and M. Lavoie. 2001. Head position as an indicator of producer and scrounger tactics in a ground-feeding bird. *Animal Behaviour* 61: 895–903.

Darwin, C. 1874. *The descent of man and selection in relation to sex*. 2nd ed. London: Murray.

David, P., T. Bjorksten, K. Fowler, and A. Pomiankowski. 2000. Condition-dependent signalling of genetic variation in stalk-eyed flies. *Nature* 406: 186–188.

Düsing, C. *1884*. *Die Regulierung des Geschlechtsverhältnisses*. Jena: Fischer.

Edwards, A.W.F. 2000. Carl Düsing (1884) on the regulation of the sex-ratio. *Theoretical Population Biology* 58: 255–257.

Fisher, R.A. 1930. *The genetical theory of natural selection*. Oxford: Clarendon Press.

Giraldeau, L.-A., and T. Caraco. 2000. *Social foraging theory*. Princeton: Princeton University Press.

Grafen, A. 1991. Biological signals as handicaps. *Journal of Theoretical Biology* 144: 517–546.

Grafen, A. 2006. A theory of Fisher's reproductive value. *Journal of Mathematical Biology* 53: 15–60.

Gross, M.R. 1985. Disruptive selection for alternative life histories in salmon. *Nature* 313: 47–48.

Gross, M.R. 1996. Alternative reproductive strategies and tactics: Diversity within sexes. *Trends in Ecology & Evolution* 11: 92–98.

Hamilton, W.D. 1964. The genetical evolution of social behaviour, I, II. *Journal of Theoretical Biology* 7: 1–52.

Hamilton, W.D. 1979. Wingless and fighting males in fig wasps and other insects. In *Reproductive competition, mate choice and sexual selection in insects*, ed. M.-S. Blum and N.A. Blum. New York: Academic Press.

Hammerstein, P. 2003. Why is reciprocity so rare in social animals? A protestant appeal. In *Genetic and cultural evolution of Cooperation*, ed. P. Hammerstein. Cambridge: MIT Press.

Hammerstein, P., and E.H. Hagen. 2005. The second wave of evolutionary economics in biology. *Trends in Ecology & Evolution* 20: 604–609.
Hofbauer, J., and K. Sigmund. 1998. *Evolutionary games and population dynamics*. Cambridge: Cambridge University Press.
Houston, A.I., and J.M. McNamara. 1999. *Models of adaptive behaviour: An approach based on state*. Cambridge: Cambridge University Press.
Maynard Smith, J. 1982. *Evolution and the theory of games*. Cambridge: Cambridge University Press.
Maynard Smith, J., and D. Harper. 2003. *Animal signals*. Oxford: Oxford University Press.
Maynard Smith, J., and G.R. Parker. 1976. The logic of asymmetric contests. *Animal Behaviour* 24: 159–175.
Maynard Smith, J., and G.R. Price. 1973. The logic of animal conflict. *Nature* 246: 15–18.
McNamara, J.M., and A.I. Houston. 1996. State-dependent life histories. *Nature* 380: 215–221.
Metz, J.A.J., R.M. Nisbet, and S.A.H. Geritz. 1992. How should we define 'fitness' for general ecological scenarios? *Trends in Ecology & Evolution* 7: 198–202.
Metz, J.A.J., S.A.H. Geritz, G. Meszéna, F.J.A. Jacobs, and J.S. van Heerwaarden. 1996. Adaptive dynamics, a geometrical study of nearly faithful reproduction. In *Stochastic and spatial structures of dynamical systems. Proceedings of the Royal Dutch Academy of Science (KNAV Verhandelingen)*, ed. S.J. van Strien and S.M. Verduyn Lunel. Amsterdam: North-Holland.
Mottley, K., and L.-A. Giraldeau. 2000. Experimental evidence that group foragers can converge on predicted producer-scrounger equilibria. *Animal Behaviour* 60: 341–350.
Noe, R., and P. Hammerstein. 1994. Biological markets: Supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology* 35: 1–11.
Nowak, M.A., and K. Sigmund. 2004. Evolutionary dynamics of biological games. *Science* 303: 793–799.
Pen, I., and F.J. Weissing. 2002. Optimal sex allocation: Steps towards a mechanistic theory. In *Sex ratios – concepts and research methods*, ed. I.C.-W. Hardy. Cambridge: Cambridge University Press.
Rice, S.H. 2004. *Evolutionary theory – mathematical and conceptual foundations*. Sunderland: Sinauer.
Shaw, R.F., and J.D. Mohler. 1953. The selective significance of the sex ratio. *American Naturalist* 87: 337–342.
Shuster, S.M., and M.J. Wade. 2003. *Mating systems and strategies*. Princeton: Princeton University Press.
Sigmund, K. 2005. John Maynard Smith and evolutionary game theory. *Theoretical Population Biology* 68: 7–10.
Spence, M. 1973. Job market signaling. *Quarterly Journal of Economics* 87: 355–374.
Spence, M. 1974. *Market signaling*. Cambridge: Harvard University Press.
Trivers, R.L. 1971. The evolution of reciprocal altruism. *Quarterly Review of Biology* 46: 35–57.
Zahavi, A. 1975. Mate selection – a selection for a handicap. *Journal of Theoretical Biology* 53: 205–214.

Game Theory in Economics, Origins of

Robert Leonard

Abstract

Game theory entered economics with the publication in 1944 of the *Theory of Games and Economic Behavior* by John von Neumann and Oskar Morgenstern. The authors were, respectively, a Hungarian mathematician and an Austrian economist. Paying attention to the scientific and cultural context, this article discusses the creation, content and impact of that work.

Keywords

Austrian economics; Cardinal utility; Coalitional game; Coalitions; Coopperative games; Courant, R; Cournot, A. A; Debreu, G; Dominance; Evolutionary biology; Existence of equilibrium; Experimental economics; Game theory; Game theory in economics, history of; General equilibrium; Hilbert, D; Mathematics and economics; Menger, K; Minimax theorem; Morgenstern, O; Nash equilibrium; Nash, J; Noncooperative games; Shapley value; Shapley. L; Side payments; Strategic equivalence; Ville, J; von Neumann, J; Weyl, H

JEL Classifications

B2

> Johnny called me; he likes my manuscript. . . . I am very happy about this. After all, it wasn't easy for me to simplify his mathematical theory, and to represent it correctly. He is working continuously without a break; it is nearly eerie. Oskar Morgenstern, Diary, 7 August 1941 (author's translation)

Thus confided Oskar Morgenstern to his wartime diary at Princeton, while working on the introductory chapter of what would become the *Theory of Games and Economic Behavior* (1944). Like

other private reflections written at the time, it speaks to the distance that lay between him, the Viennese economist, and 'Johnny' von Neumann, the Hungarian mathematician. The two exiles had come to Princeton by quite different paths, and, in 1941, had known each other well for only 2 years. A product of the Austrian school of economics, Morgenstern had strong critical faculties and epistemological interests, but limited mathematical training. Von Neumann, on the other hand, was a Hungarian mathematician of first rank, with little time for philosophical speculation but boundless confidence in the application of mathematics across the scientific domain. Yet, differences notwithstanding, they managed to forge a fruitful partnership, writing a landmark 600-page book on mathematical social science that would mark the creation of game theory and link them thereafter in the public eye.

Von Neumann (1903–1957), the privately tutored mathematical prodigy, came from a prosperous banking family of assimilated Hungarian Jews. It is not insignificant for the present subject that, during his formative years, he was witness to political upheaval, including not only the First World War but also, in Hungary, the 1919 Communist revolution of Bela Kun and its subsequent brutal suppression. Indeed, the Kun regime saw the von Neumann family temporarily flee Budapest. He also watched the growth of anti-semitism in Hungary, which would increasingly restrict the opportunities available to even well-integrated Jews such as himself. In the mid-1920s, he completed degrees in mathematics and chemical engineering at Budapest and Zurich, during that time writing several papers, mainly in the areas of axiomatic set theory and the consistency of mathematics. In 1926, he became postdoctoral fellow at the University of Göttingen, then a world centre in mathematics, whose rich environment allowed him to work close to not only its leader David Hilbert but other luminaries such as Richard Courant and Hermann Weyl. During this period, he continued working on set theory and foundations and, in particular, the mathematical theory of quantum mechanics (see von Neumann 1932). In these works, there emerge features that would characterize von Neumann's use of mathematics in social science, including an emphasis on achieving axiomatic description of the field under study and, perhaps inherited from quantum mechanics, a belief in the inherently probabilistic nature of the world.

Both of these features surfaced in another of von Neumann's Göttingen papers, one which stood apart from his main interests. This was his 1928 'Zur Theorie der Gesellschaftsspiele', the theory of parlour games, first presented when von Neumann was 23 years old. Were space restrictions unimportant here, we could explore the rich background to the mathematical treatment of games. Chess held a great place in the Jewish culture of *Mitteleuropa* at the turn of the twentieth century: from the psychological investigation of the thought processes of the grandmasters to the use of the game as source of inspiration in novelists' fiction. Legendary chess champion and mathematician, Emanuel Lasker, drew on the game for inspiration as he wrote about the workings of social life. Other mathematician contemporaries of his wondered whether so human an activity as chess could be made amenable to formal treatment. The key figures here were Ernest Zermelo in 1913 and then, in the 1920s, von Neumann's Hungarian contemporaries, Dénes König and László Kalmár. Independently, in Paris, again in the 1920s, French mathematician, Emile Borel, drew on his experience as a player of cards rather than chess to begin constructing a mathematical analysis of games involving strategy and to probe the question of equilibrium play. Von Neumann's paper may be regarded as the crowning contribution of these mathematical investigations.

His paper is a brilliant description of the generic two-person, zero-sum game, that is, in which the interests of the two players are directly opposed. He defines the game by the strategies available to both players and their associated payoffs, and, never too concerned about elegance in mathematics, gives a tortuously difficult proof of the existence of a minimax equilibrium. This is a preferred way to play, possibly requiring that strategies be chosen in a probabilistic manner, that allows each player to minimize the amount ceded to the other. The paper, which brought the discussion of the existence of such an equilibrium

to a close, finished with preliminary suggestions about how to extend the analysis to games of three or more players, in terms of coalitions and their winnings. But it went no further than that, and neither did von Neumann. Apart from some unpublished work at the time, in which he showed how the strategy of 'bluffing' in a simple two-person poker corresponded to the mathematically rational way to play, he essentially put the theory of games aside. Following a period as *Privatdozent* at the University of Berlin, he spent 6 months of 1930 at Princeton University's Department of Mathematics. The following year, gauging that his opportunities in Europe were limited, he moved permanently to the United States, where, along with Albert Einstein, he became one of the first members of the newly founded Institute for Advanced Study, close to Princeton University.

Morgenstern (1906–1976) was part of the thriving interwar economics community in Vienna constituted by such figures as Ludwig von Mises, Hans Mayer, Friedrich Hayek, Fritz Machlup and Gottfried Haberler. Having obtained his *Habilitation* in 1928, he became lecturer at the University of Vienna and then director of the Rockefeller-financed *Institut für Konjunkturforschung* (Business Cycle Institute). Although an Austrian economist, he was not as ardent an advocate of laissez-faire liberalism as Mises and Hayek, being closer to von Wieser and Hans Mayer, both of whom were more accepting of public intervention and strong government. As Institute director, Morgenstern also had to accommodate himself to the authoritarian Christian Social government that governed Austria between 1934 and 1938.

Like his fellow members of the Austrian School, not least Hayek, Morgenstern was critical of general equilibrium theory, and particularly of what the Viennese viewed as its lack of precision in treating the knowledge, beliefs and expectations of forward-looking economic actors (see Morgenstern 1935). Unlike many of his Viennese colleagues, however, and notwithstanding his own limited training in the subject, Morgenstern learned to see the further application of mathematics as a means by which to improve the rigour of economic theory. In this, he differed from Mises and Hayek, for example, who were quite sceptical about what was to be gained by applying mathematics to the non-mechanical realm of human action. Here, Morgenstern was influenced by his contact with mathematician Karl Menger, who was son of the founder of the Austrian School of economics, was close to the Vienna Circle, and was leader of that small coterie of mathematical economists in interwar Vienna, which included Karl Schlesinger, Abraham Wald and a young Franz Alt. Menger's friendship, activities and writings, including his 1934 book on ethics and social compatibility, were of fundamental importance to Morgenstern at this time (see Menger 1934; Leonard 1998).

Like von Neumann, Morgenstern's career was shaped in part by social and political upheaval. In 1927, Vienna was the theatre of political violence between the Austrian Right and the Socialists. In 1934, there was a civil war in Austria, as the conservative Chancellor Dollfuss crushed the Left, and in 1938 the country was annexed by Germany. This marked the demise of one of the most intellectually and culturally active cities of interwar Europe. Although not Jewish, Morgenstern found himself ousted from his Institute and, leaving Austria in 1938, he took a position at the department of economics at Princeton University. The latter was then very much a sleepy gentlemen's college, so that, for sophisticated intellectual company, Morgenstern found himself turning towards the mathematicians and physicists at the Institute of Advanced Study.

By this time, von Neumann was already returning to game theory. Throughout the 1930s his correspondence shows him to be a very astute observer from afar of the political situation in Europe, and it was against this background of irrationality and social instability that he returned, at the close of the decade, to the development of a mathematics of alliances and coalitions. In late 1940 and 1941, quite independent of Morgenstern, he extended his 1928 theory to the treatment of three, four and more players, culminating in an analysis of the general n-person, non-zero-sum game. These ideas on games attracted Morgenstern, who saw in them not so

much a theory of the social order as a response to his Viennese concerns about how to model the interaction between economic agents. The ensuing collaboration with von Neumann in the period 1941–1943 resulted in the *Theory of Games and Economic Behavior*, the entire technical apparatus of which was provided by von Neumann, and the introduction and general orientation by Morgenstern (see Leonard 1995, 2008).

Contents of *Theory of Games and Economic Behavior*

That introductory chapter is the most accessible, and therefore most widely read and cited, part of the *Theory of Games*. It is at once a defence of the use of mathematics in social science and a critique of the prevailing state of mathematical economics. Von Neumann's influence, and almost religious faith in the supremacy of mathematical formalism, is clear throughout. There is nothing intrinsically different about social science, it is claimed, that renders it inaccessible to mathematical treatment. Natural phenomena, whether or not they concern human behaviour, are potential repositories of mathematics, the richness of which is likely to be correlated with the empirical prominence of the field. Social and economic activity is of such great worldly importance that it is likely to, so to speak, generate a mathematics of its own.

The most prominent treatment of the area, general equilibrium theory, is merely the imitative grafting of physical science methods onto social science. This brings with it assumptions about underlying continuity of change, whereas the social domain likely requires attention to discretely separate structures, and thus the use of a different mathematics. General equilibrium theory has also failed to account for the properly interactive nature of social behaviour, particularly that which is manifest in situations involving 'small' numbers of agents, be they involved in the exchange of goods or in the distribution of gains through the formation of social and political groups. Throughout the book, von Neumann's preference for 'modern', discrete mathematics (that is, set theory and combinatorics) over the differential and integral calculus is evident. Several pages are devoted to defending the use of cardinal, or numerical, utilities, with the axiomatic proof of the existence of a cardinal utility function being included in an Appendix to the second edition, published in 1947.

Chapter 2 lays out the notion of a game, introducing the mathematical concepts of sets and partitions, and showing how the game may be described axiomatically in these terms. The whole is presented as a piece of modern, axiomatic mathematics in the spirit of Hilbert, which is to say that, although the axioms are stimulated by the common-sense features of games, the latter are soon let recede into the background and the theory pursued in a spirit of relative abstraction. While the mathematics is being followed through, the empirical is held at arm's length and everyday terms are introduced in inverted commas – hence, 'class', 'discrimination', 'exploitation', and so on. Only during periodic returns to the heuristics is the vocabulary of the everyday re-invoked, and the 'common sense' meaning of the results discussed. The minimax theorem is proved in the next chapter, using, not von Neumann's earlier proof, but a modification of the elementary 1938 proof by Borel's student Jean Ville, based on the theory of convex sets. From here on, chapter by chapter, von Neumann systematically goes through the zero-sum game for three, four and more players, exploring their combinatorial possibilities for coalition-formation and compensations (side payments). Each game is described in terms of its characteristic function, which shows the maximal payoff available to each possible coalition of the game, assuming that the coalition plays minimax against its complement and that utility is transferable between players. In Chap. 9, the concept of strategic equivalence is introduced to show how the move from the zero-sum restriction to a constant sum retains the basic features of the game, thus allowing it to be solved by the same means. In the eleventh chapter, von Neumann drops the zero (or constant) sum restriction, moving to the 'general game'.

The central theoretical part of the *Theory of Games* is von Neumann's solution to coalitional games, the stable set. The solution is a

'complicated combinatorial catalogue', indicating the minimum each participant can get if he behaves rationally. He may, of course, get more if the others behave 'irrationally', that is, make mistakes. Were the solution to consist of a single imputation – a vector of amounts to be received by each player – then the 'structure of the society under consideration would be extremely simple: There would exist an absolute state of equilibrium in which the quantitative share of every participant would be precisely determined' (1944, p. 34). However, such a unique solution does not generally exist – a given society can be organized in various ways – so the notion needs to be broadened. The solution is thus a *set* of possible imputations.

> Any particular alliance describes only one particular consideration which enters the minds of the participants when they plan their behavior. Even if a particular alliance is ultimately formed, the division of the proceeds between the allies will be decisively infuenced by the other alliances which each one might alternatively have entered . . . It is, indeed, this whole which is the really significant entity, more so than its constituent imputations. Even if one of these is actually applied, i.e., if one particular alliance is actually formed, the others are present in a 'virtual' existence: Although they have not materialized, they have contributed essentially to shaping and determining the actual reality. (1944, p. 36)

In an n-person game, therefore, a 'solution should be a system of imputations possessing in its entirety some kind of balance and stability the nature of which we shall try to determine. We emphasize that this stability – whatever it may turn out to be – will be a property of the system as a whole and not of the single imputations of which it is composed' (p. 36).

This stability is based on the notion of 'domination'. One imputation, x, is said to dominate another, y, 'when there exists a group of participants each one of which prefers his individual situation in x to that in y, and who are convinced that they are able, as a group – i.e. as an alliance – to enforce their preferences' (p. 38). Dominance, which is not a transitive ordering since the demurring coalition may be different in each case, forms the basis for game solutions. Von Neumann defines the solution to an n-person game as a set of imputations, S, with the following characteristics:

> No imputation y contained in S is dominated by an imputation x contained in S.
> Every y not contained in S is dominated by some x contained in S.

A solution is thus not a single imputation but a set of possible imputations, and such a set is stable in so far as its member imputations do not dominate each other and every imputation outside the set is dominated by at least one imputation inside. Further, not only is a solution comprised of possibly many imputations, linked by these stability criteria, but a given game may have many solutions. To take a simple example, consider the zero–sum game in which a 'pie' of value 1 has to be divided amongst three people. It has the following four solutions:

1. (1/2, 1/2, 0) (1/2, 0, 1/2) (0, 1/2, 1/2)
2. $(x_1, \quad x_2, \quad c)$ where $0 \leq c \leq 1/2$ and $x_1 + x_2 + c = 1$
3. $(c, \quad x_2, \quad x_3)$ where $0 \leq c \leq 1/2$ and $x_2 + x_3 + c = 1$
4. $(x_1, \quad c, \quad x_3)$ where $0 \leq c \leq 1/2$ and $x_1 + x_3 + c = 1$

Here, not only are there multiple solutions, but three of those actually admit an infinite number of possible imputations. Note also that the observation of a given imputation, such as (1/2, 1/2, 0), says little about which solution obtains, as that imputation could occur in any of the four solutions above.

The question of which solution will obtain in a given situation, the authors say, can be broached only by considering 'standards of behaviour', the various rules, customs or institutions governing social organization at the time. These are extra-game considerations, not contained in the information provided by the characteristic function. To understand the analogy, von Neumann and Morgenstern advise the reader to 'temporarily forget the analogy with games and think entirely in terms of social organization' (p. 41, n. 1):

> Let the physical basis of a social economy be given, – or to take a broader view of the matter, of a society. According to all tradition and

> experience human beings have a characteristic way of adjusting themselves to such a background. This consists of not setting up one rigid system of apportionment, i.e. of imputation, but rather a variety of alternatives, which will probably express some general principles but nevertheless differ among themselves in many particular respects. This system of imputations describes the 'established order of society' or 'accepted standard of behavior'.

Thus, in the above game, in solution 2, player 3 is held to an amount, c, that may be as small as zero, or as high as 1/2. The actual value of c would reflect the norms governing that player's social standing. Depending on tradition, the marginal member might or might not be completely exploited. As von Neumann and Morgenstern write: 'A theory which is consistent at this point cannot fail to give a precise account of the entire interplay of economic interest, influence and power' (p. 43).

When one considers the personal context in which von Neumann developed this social theory, his letters of the time dwelling on European instability, the strategic alliances involving the Germans and the Allies, the plight of the Hungarian Jews, which directly affected his family, and one then reads the *Theory of Games*, with its emphasis on stability and its pervasive reference to norms, discrimination and power, the book appears as his attempt, not simply to replace general equilibrium theory, but to achieve a mathematical description of social organization, broadly defined. And, to the end of his life, von Neumann spoke of it in these terms. For example, in 1955, at a Princeton conference on game theory, when mathematician John Nash raised the problem of the great multiplicity of solutions to cooperative games, von Neumann replied 'that this result was not surprising in view of the correspondingly enormous variety of observed stable social structures; many differing conventions can endure, existing today for no better reason than that they were here yesterday' (Wolfe 1955, p. 25).

Role and Impact of the Book

The initial influence of the *Theory of Games* was felt, however, not in the area of economic or social theory per se, but in that of military strategy. During the early 1940s, while Morgenstern remained at Princeton struggling with the technical draft chapters, von Neumann was out in the world, increasingly heavily involved as mathematical advisor to various branches of the American military forces. Through his influence on the work of mathematicians at the Princeton branch of the Statistical Research Group and at the Anti-submarine Warfare Operations Research Group, game theory became an element in mathematical models of military engagements such as submarine-search and bombing strategy. Such military models involved the application of a very small part of the mathematics – usually centred on minimax theorem – to specific, confined problems. Thus, game theory qua operations research was far removed from the ambitious, abstract representation of the social order that von Neumann had pursued in the *Theory of Games*. Be that as it may, it was the perceived success of operations research during the Second World War that provided the impetus for the Army Air Corps' creation of the RAND Corporation in the late 1940s, where models of this kind continued to be developed, and for the next decade game theory was given strong institutional support. As it happens, there is little evidence that, throughout the 1950s at RAND, these game theoretic models were of anything other than very limited influence in quantitatively shaping particular strategic decisions. It is incontestable, however, that the language, terminology and 'thought framework' of game theory became important to the strategic mindset that dominated the Cold War period, helping shape such books as Herman Kahn's *Thinking the Unthinkable* and Thomas Schelling's *The Strategy of Conflict*.

The *Theory of Games* also set new standards for mathematical rigour in the field of economic theory. For example, before leaving France to move to the United States, economic theorist Gerard Debreu read the book in Salzburg, Austria, at a summer-school run by Harvard University. Though Debreu would never work on game theory, the book shaped his thinking greatly. His pathbreaking *Theory of Value* (1959), an axiomatic treatment of Walrasian general equilibrium theory, refers to the outstanding influence of von Neumann and Morgenstern (1944) 'which freed

mathematical economics from its traditions of differential calculus and compromises with logic' (1959, p. x). Debreu's stance, too, on the relationship between the mathematics of general equilibrium and the empirical economic substrate is exactly that of von Neumann on games: 'the theory, in the strict sense, is logically entirely disconnected from its interpretations' (p. x). This austere, formal view shaped an entire generation of economic 'high theorists' from the 1950s till the 1980s, during which general equilibrium theory was the pinnacle of intellectual achievement in the discipline. That von Neumann's game theory should have ended up providing sustenance for Walrasian general equilibrium is one of the many historical ironies in the intellectual history of our field.

It was also in the post-war military–academic milieu that a new generation of game theorists proper came of age. Whether at Princeton or the RAND Corporation, or alternating between the two, young mathematicians such as Lloyd Shapley and John Nash took game theory and made it their own. Shapley, a towering influence in the game theory community from the 1950s onwards, produced, amongst other things, the Shapley value, which described the solution to a coalitional game in terms of the amount brought by each player to an average, randomly formed coalition. For his Ph.D. thesis, Nash sought to provide for n-person games a solution that was as well-defined as von Neumann's minimax for the two-person game. He made a conceptual division of games into cooperative, in which coalitions are permitted, and non-cooperative, in which players act in isolation. For the latter, he proved the existence, under specific conditions, of what he called an 'equilibrium point', later the Nash equilibrium (see Nash 1950a, b, 1951). That von Neumann found this non-cooperative approach to be rather trivial is understandable in the light of the ambitious social theory he was pursuing, but that leaves unchanged the fact of his influence. It was also at this time that the work of Augustin Cournot was redisovered and reinterpreted in the light of the Nash equilibrium (see Leonard 1994; Dimand and Dimand 1996).

Subsequent work on non-cooperative game theory by Harsanyi, Selten, Aumann, Kreps and others resulted in a veritable transformation of the microeconomic canon and shaped modelling in industrial organization, international trade and a range of areas (see Dimand 2000). The field of experimental economics, currently in rapid expansion, owes it existence in part to the appearance of game theory. Although von Neumann himself voiced scepticism as to the ability of laboratory experimentation to shed light on the stable set, game theory did provide a structured basis for empirically testing the theory of individual decision, via its utility axioms, and various solution concepts, both cooperative and non-cooperative. This experimentation, too, began at the RAND Corporation (see Kalisch et al. 1954). It should also be mentioned that, under the influence on John Maynard Smith, the theory of games has had a great impact on the field of evolutionary biology (see Maynard Smith 1988).

In short, although it quickly attained the status of a classic, which is to say that it was cited by many but read by few, the *Theory of Games and Economic Behavior* set in motion developments that, in ways sometimes quite unintended by the its authors, gradually reshaped the warp and weft of the economics discipline. From the recasting of the economic agent as a strategic player to the reshaping of entire fields of economics; from the introduction to general equilibrium and social welfare theory of axiomatic methods and discrete mathematics to the rise of experimental economics, the direct and indirect effects of von Neumann and Morgenstern's wartime book have been profound and long-lasting.

See Also

- ▶ game theory
- ▶ Morgenstern, Oskar (1902–1977)
- ▶ Nash, John Forbes (born 1928)
- ▶ von Neumann, John (1903–1957)

Bibliography

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

Dimand, R. 2000. Strategic games: From theory to application. In *The history of applied economics*, ed. R.-E. Backhouse and J. Biddle. Annual Supplement to *History of Political Economy* 32, 199–225.

Dimand, M.A., and R.W. Dimand. 1996. *A history of game theory, Vol. I. From the beginnings to 1945*. London/New York: Routledge.

Kahn, H. 1962. *Thinking about the unthinkable*. New York: Horizon Press.

Kalisch, G., J.W. Milnor, J. Nash, and E.D. Nering. 1954. Some experimental n-person games. In *Decision processes*, ed. R.M. Thrall, C.H. Coombs, and R.L. Davis. New York: Wiley.

Leonard, R. 1994. Reading Cournot, reading Nash: The creation and stabilisation of the Nash equilibrium. *Economic Journal* 104: 492–511.

Leonard, R. 1995. From parlor games to social science: Von Neumann, Morgenstern, and the creation of game theory, 1928–1944. *Journal of Economic Literature* 33: 730–761.

Leonard, R. 1998. Ethics and the excluded middle: Karl Menger and social science in interwar Vienna. *Isis* 89(1): 1–26.

Leonard, R. 2008. *Von Neumann, Morgenstern and the creation of game theory, 1900–1960*. New York/Cambridge: Cambridge University Press.

Maynard Smith, J. 1988. *Games, sex and evolution*. New York/Toronto: Harvester-Wheatsheaf.

Menger, K. 1934. *Moral, Wille und Weltgestaltung. Grundlegung zur Logik der Sitten*. Vienna: Julius Springer. Trans. as Morality, decision and social organization: Towards a logic of ethics. Dordrecht: Reidel, 1974.

Morgenstern, O. 1935. Vollkommene Voraussicht und wirtschaftliches Gleichgewicht. *Zeitschrift für Nationalökonomie* 6, 337–357 Trans. F. Knight, mimeo, University of Chicago. Repr. in *Selected economic writings of Oskar Morgenstern*, ed. A. Schotter. New York: NYU Press, 1976.

Morgenstern, O. *Diary*. Oskar Morgenstern Papers, Special Collections Library, Duke University, USA.

Nash, J. 1950a. Equilibrium points in N-person games. *Proceedings of the National Academy of Science* 36, 48–49.

Nash, J. 1950b. Non-cooperative games. Ph.D. thesis, Princeton University.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* 54: 286–95.

Schelling, T. 1960. *The strategy of conflict*. Cambridge, MA: Harvard University Press.

von Neumann, J. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100, 295–320 Trans. S. Bargmann as 'On the theory of games of strategy', in *Contributions to the theory of games*, vol. 4, ed. A.W. Tucker and R.D. Luce. Princeton: Princeton University Press, 1959.

von Neumann, J. 1932. *Mathematische Grundlagen der Quantenmechanik*. Berlin: J. Springer. Trans. R. Beyer as *Mathematical foundations of quantum mechanics*. Princeton: Princeton University Press, 1955.

von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.

von Neumann, J., and O. Morgenstern. 1947. *Theory of games and economic behavior*, 2nd ed. Princeton: Princeton University Press.

Wolfe, P. (ed). 1955. Report of an informal conference on recent developments in the theory of games. Mimeo, Logistics Research Project, Department of Mathematics, Princeton University, Princeton.

Games in Coalitional Form

Ehud Kalai

Abstract

How should a coalition of cooperating players allocate payoffs to its members? This question arises in a broad range of situations and evokes an equally broad range of issues. For example, it raises technical issues in accounting, if the players are divisions of a corporation, but involves issues of social justice when the context is how people behave in society.

Despite the breadth of possible applications, coalitional game theory offers a unified framework and solutions for addressing such qsts. This article presents some of its major models and proposed solutions.

Keywords

Additive games; Auction game; Aumann–Shapley prices; Axiomatic characterizations; Balanced games; Bargaining; Bargaining sets; Coalition; Coalition formation; Coalitional game; Coalitional monotonicity; Communication graph; Consistency; Cooperation; Cooperative game theory; Core; Cost allocation; Dummy player; Egalitarian solution; Egalitarian value; Equivalence th; Flow game; Folk th; Game theory; Games in coalitional form; Grand coalition; Harsanyi value; Imputation; Independence of irrelevant alternatives; Individual monotonicity; Individual rationality; Kalai–Smorodinsky solution; Kernel; Large market games; Law of large numbers; Linear programming game; Majority

game; Market games; Maschler–Perles solution; Matching; Matching and market design; Middleman; Monotonicity; Nash bargaining games; Nash solution; Network games; No transferable-utility game; Nucleolus; Pareto optimality; Partition games; Population monotonicity; Profit sharing; Raiffa solution; Shapley value; Simple games; Spanning-tree game; Stable sets; Strategic game theory; Supperadditive games; Transferable-utility game; Voting games

JEL Classifications
C7

Introduction

In their seminal book, von Neumann and Morgenstern (1944) introduced two theories of games: strategic and coalitional. Strategic game theory concentrates on the selection of strategies by payoff-maximizing players. coalitional game theory concentrates on coalition formation and the distribution of payoffs.

The next two examples illustrate situations in the domain of the coalitional approach.

Games with No Strategic Structure

Example 1 Cost Allocation of a Shared Facility *Three municipalities, E, W, and S, need to construct water purification facilities. Costs of individual and joint facilities are described by the cost function c: c(E) = 20, c(W) = 30, and c(S) = 50; c(E,W) = 40, c(E,S) = 60, and c(W,S) = 80; c (E,W,S) = 80. For example, a facility that serves the needs of W and S would cost $80 million.The optimal solution is to build, at the cost of 80, one facility that serves all three municipalities. How should its cost be allocated*?

Games with Many Nash Equilibria

Example 2 Repeated Sales *A seller and a buyer play the following stage game on a daily basis. The seller decides on the quality level, H, M, or L, of the item sold (at a fixed price); without knowledge of the seller's selected quality, the buyer decides whether or not to buy. If she does not buy, the payoffs of both are zero; if she buys, the corresponding payoffs are (0, 3), (3, 2) or (4, 0), depending on whether the quality is H, M, or L. Under perfect monitoring of past choices and low discounting of future payoffs, the folk theorem of repeated games states that any pair of numbers in the convex hull of (0, 0),(0, 3), (3, 2), and (4, 0) are Nash-equilibrium average payoffs. What equilibrium and what average payoffs should they select*?

We proceed with a short survey of the major models and selected solution concepts. More elaborate overviews are available in Game Theory, Myerson (1991), and other surveys mentioned below.

Types of Coalitional Game

In what follows, N is a fixed set of *n players*; the set of *coalitions* C consists of the nonempty subsets of N; $|S|$ denotes the number of players in a coalition S. The terms 'profile' and 'S-profile' denote vectors of items (payoffs, costs, commodities, and so on) indexed by the names of the players.

For every coalition S, R^S denotes the $|S|$-dimensional Euclidean space indexed by the names of the players; for single-player coalitions the symbol i replaces fig. A profile $u^S \in R^S$ denotes *payoffs* u_i^S of the players $i \in S$.

Definition 1 *An (n person) game (also known as a game with no transferable utility, or NTU game) is a function V that assigns every coalition S a set* $V(S) \subset R^s$.

Remark 1 *The initial models of coalitional games were presented in von Neumann and Morgenstern (1944) for the special case of TU games described below, Nash (1950) for the special case of two-person games, and Aumann and Peleg (1960) for the general case.*

G

The interpretation is that $V(S)$ describes all the *feasible payoff profiles* that the coalition S can generate for its members. Under the assumption that the *grand coalition* N is formed, the central question is which payoff profile $u^N \in V(N)$ to select. Two major considerations come into play: the relative strength of different coalitions, and the relative strength of players within coalitions.

To separate these two issues, game theorists study the two simpler types of games defined below: TU games and bargaining games. In TU games the players in every coalition are symmetric, so only the relative strength of coalitions matters. In bargaining games only one coalition is active, so only the relative strength of players' within that coalition matters. Historically, solutions of games have been developed first for these simpler classes of games, and only then extended to general (NTU) games. For this reason, the literature on these simpler classes is substantially richer than the general theory of (NTU) games.

Definition 2 *V is a transferable-utility game (TU game) if for a real-valued function $v = (v(S))_{S \in \ell}$, $V(S) = \{u^S \in R^S : \sum_i u_i^S \leq v(S)\}$.*

It is customary to identify a TU game by the function v instead of V.

TU games describe many interactive environments. Consider, for example, any environment with individual outcomes consisting of prizes p and monetary payoffs m, and individual utilities that are additive and separable in money ($u_i(p, m) = v_i(p) + m$). Under the assumption that the players have enough funds to make transfers, the TU formulation presents an accurate description of the situation.

Definition 3 *A Nash (1950) bargaining game is a two-person game. An n-person bargaining game is a game V in which $V(S) = \times_{i \in S} V(i)$ for every coalition $S \subsetneq N$.*

Remark 2 Partition games (Lucas and Thrall 1963) *use a more sophisticated function V to describe coalitional payoffs. For every partition of the set of players $\pi = (T_1, T_2, \ldots, T_m)$, $V_\pi(T_j)$ is the set of T_j's feasible payoff profiles, under the cooperation structure described by π. Thus, what is feasible for a coalition may depend on the strategic alignment of the opponents. The literature on partition games is not highly developed.*

Some Special Families of Games

Coalitional game theory is useful for analysing special types of interactive environments. And conversely, such special environments serve as a laboratory to test the usefulness of game theoretic solutions. The following are a few examples.

Profit Sharing and Cost Allocation

Consider a partnership that needs to distribute its total profits, $v(N)$, to its n individual partners. A profit-distribution formula should consider the potential profits $v(S)$ that coalitions of partners S can generate on their own. A TU game is a natural description of the situation.

A cost allocation problem, like Example 1, can be turned into a natural TU game by defining the worth of a coalition to be the savings obtained by joining forces: $v(S) = \sum_{i \in S} c(i) - c(S)$.

Examples of papers on cost allocation are Shubik (1962) and Billera et al. (1978). See Young (1994) for an extensive survey.

Markets and Auctions

Restricting this discussion to simple exchange, consider an environment with n traders and m commodities. Each trader i starts with an initial *bundle* ω_i^0, an m-dimensional vector that describes the quantities of each commodity he owns. The utility of player i for a bundle ω_i is described by $u_i(\omega_i)$. An S-profile of bundles $\omega = (\omega_i)_{i \in S}$ is feasible for the coalition S if $\sum_{i \in S} \omega_i = \sum_{i \in S} \omega_i^0$.

Definition 4 *A game V is a market game, if for such an exchange environment (with assumed free-disposal of utility),*

$$V(S) = \{u^S \in R^S : \text{for some } S - \text{feasible profile of bundles } \omega,\ u_i^S \leq u_i(\omega_i) \text{ for every } i \in S.$$

Under the assumptions discussed earlier (additively separable utility and sufficient funds)

the market game has the more compact TU description: $v(S) = max_{\omega} \sum_{i \in S} u_i(\omega_i)$, with the max taken over all S-feasible profiles ω.

As discussed below, market games play a central role in several areas of game theory.

Definition 5 *An auction game is a market game with a seller whose initial bundle consists of items to be sold, and bidders whose initial bundles consist of money.*

Matching Games

Many theoretical and empirical studies are devoted to the subject of efficient and stable matching: husbands with wives, sellers with buyers, students with schools, donors with receivers, and more; see matching and market design. The first of these was introduced by Gale and Shapley in their pioneering study (1962) using the following example.

Consider a matching environment with q males and q females. Payoff functions $u_m(f)$ and $u_m(none)$ describe the utilities of male m paired with female f or with no one; $u_f(m)$ and $u_f(none)$ describe the corresponding utilities of the females. A pairing PS of a coalition S is a specification of male-female pairs from S, with the remaining S members being unpaired.

Definition 6 *A game V is a marriage game if for such an environment, $V(S) = \left\{u^S \in R^S : \text{for some pairing } P_S, u_i^S \leq u_i(P_S) \text{ for every } i \in S\right\}$.*

Solutions of marriage games that are efficient and stable (that is, no divorce) can be computed by Gale–Shapley algorithms.

Optimization Games

Optimization problems from operations research have natural extensions to multiperson coalitional games, as the following examples illustrate.

Spanning-Tree Games

A cost-allocation TU spanning-tree game (Bird 1976) is described by an undirected connected graph, with one node designated as the centre C and every other node corresponding to a player. Every arc has an associated nonnegative connectivity cost. The cost of a coalition S, $c(S)$, is defined to be the minimum sum of all the arc costs, taken over all subgraphs that connect all the members of S to C.

Flow Games

A TU flow game (Kalai and Zemel 1982b) is described by a directed graph, with two nodes, s and t, designated as the *source* and the *sink*, respectively. Every arc has an associated capacity and is owned by one of the n players. For every coalition S, $v(S)$ is the maximal s-to-t flow that the coalition S can generate through the arcs owned by its members.

Linear Programming Games

Finding minimal-cost spanning trees and maximum flow can be described as special types of linear programmes. Linear (and nonlinear) programming problems have been generalized to multiperson games (see Owen 1975; Kalai and Zemel 1982a; Dubey and Shapley 1984). The following is a simple example.

Fix a $p \times q$ matrix A and a q-dimensional vector w, to consider standard linear programmes of the form max wx s.t. $Ax \leq b$. Endow each player i with a p-dimensional vector b_i, and define the linear-programming TU game v by $v(S) = max_x wx \ \ s.t. Ax \leq \sum_{i \in S} b_i$.

Simple Games and Voting Games

A TU game is simple if for every coalition S, $v(S)$ is either zero or 1. Simple games are useful for describing the power of coalitions in political applications. For example, if every player is a party in a certain parliament, then $v(S) = 1$ means that under the parliamentary rules the parties in the coalition S have the ability to pass legislation (or *win*) regardless of the positions of the parties not in S; $v(S) = 0$ (or S *loses*) otherwise.

In applications like the one above, just formulating the game may already offer useful insights into the power structure. For example, consider a parliament that requires 50 votes in order to pass legislation, with three parties that have 12 votes, 38 votes, and 49 votes, respectively. Even though the third party seems strongest, a simple formulation of the game yields the symmetric *three-*

person majority game: any coalition with two or more parties wins; single-party coalitions lose.

Beyond the initial stage of formulation, standard solutions of game theory offer useful insights into the power structure of such institutions and other political structures (see, for example, Shapley and Shubik 1954; Riker and Shapley 1968; Brams et al. 1983).

Solution Concepts

When cooperation is beneficial, which coalitions will form and how would coalitions allocate payoffs to their members? Given the breadth of situations for which this question is relevant, game theory offers several different solutions that are motivated by different criteria. In this brief survey, we concentrate on the Core and on the Shapley value.

Under the assumptions that utility functions can be rescaled, that lotteries over outcomes can be performed, and that utility can be freely disposed of, we restrict the discussion to games V with the following properties.

Every $V(S)$ is a compact convex subset of the nonnegative orthant R^S_+, and it satisfies the following property: if $w^S \in R^S_+$ with $w^S \leq u^S$ for some $u^S \in V(S)$, then $w^S \in V(S)$. And for single player coalitions, assume $V(i) = \{0\}$. For TU games this means that every $v(S) \geq 0$, the corresponding $V(S) = \left\{u^S \in R^S_+ : \sum_{i \in S} u^S_i \leq v(S)\right.$, and for each i, $v(i) = 0$.

In addition, we assume that the games are *superadditive*: for any pair of disjoint coalitions T and S, $V(T \cup S) \supseteq V(T) \times V(S)$; for TU games this translates to $v(T \cup S) \geq v(T) + v(S)$. Under superadditivity, the maximal possible payoffs are generated by the grand coalition N. Thus, the discussion turns to how the payoffs of the grand coalition should be allocated, ignoring the question of which coalitions would form.

A payoff profile $u \in R^N$ is *feasible for a coalition* S, if $u^S \in V(S)$, where u^S is the projection of u to R^S. The translation to TU games is that $u(S) \equiv \sum_{i \in S} u_i \leq v(S)$. A profile $u \in R^N$ can be *improved upon* by the coalition S if there is an S-feasible profile w with $w_i > u_i$ for all $i \in S$.

Definition 7 *An imputation of a game is a grand-coalition-feasible payoff profile that is both individually rational (that is, no individual player can improve upon it) and Pareto optimal (that is, the grand coalition cannot improve upon it).*

Given the uncontroversial nature of individual rationality and Pareto optimality, solutions of a game are restricted to the selection of imputations.

The Core

Definition 8 *The* core *of a game* (*see Shapley* 1952*, and Gillies* 1953*, for TU, and Aumann* 1961*, for NTU*) is the set of imputations that cannot be improved upon by any coalition.

The core turns out to be a compact set of imputations that may be empty. In the case of TU games it is a convex set, but in general games (NTU) it may even be a disconnected set. The core induces stable cooperation in the grand coalition because no sub-coalition of players can reach a consensus to break away when a payoff profile is in the core.

Remark 3 *More refined notions of stability give rise to alternative solution concepts, such as the stable sets of von Neumann and Morgenstern (*1944*), and the kernel and bargaining sets of Davis and Maschler (*1965*). The nucleolus of Schmeidler (*1969*), with its NTU extension in Kalai (*1975*), offers a 'refinement' of the core. It consists of a finite number of points (exactly one for TU games) and belongs to the core when the core is not empty. For more on these solutions, see Maschler (*1992*) and game theory.*

Unfortunately, games with an empty core are not unusual. Even the simple three-person majority game described in section "Simple Games and Voting Games" has an empty core (since among any three numbers that sum to one there must be a pair that sums to less than one, there are always two players who can improve their payoffs).

TU Games with Nonempty Cores

Given the coalitional stability obtained under payoff profiles in the core, it is desirable to know in which games the core is nonempty.

Bondareva (1963) and Shapley (1967) consider 'part-time coalitions' that meet the availability constraints of their members. In this sense, a collection of nonnegative coalitional weights $\lambda = (\lambda_S)_{S \in \ell}$ is *balanced*, if for every player i, $\sum_{S\,:\,i \in S} \lambda_S = 1$. They show that a game has a nonempty core if and only if the game is *balanced*: for every balanced collection λ, $\sum_S \lambda_S v(S) \leq v(N)$.

As Scarf (1967) demonstrates, all market games have nonempty cores and even the stronger property of having nonempty subcores: for every coalition S, consider the subgame v_S which is restricted to the players of S and their subcoalitions. The game v *has nonempty subcores*, if all its subgames v_S have nonempty cores.

By applying the balancedness condition repeatedly, one concludes that a game has nonempty subcores if and only if the balancedness condition holds for all its subgames v_S. Games with this property are called *totally balanced*.

Since Shapley and Shubik (1969a) demonstrate the converse of Scarf's result, a game is thus totally balanced if and only if it is a market game. Interestingly, the following description offers yet a different characterization of this family of games.

A game w is *additive* if there is a profile $u \in R^N$ such that for every coalition S, $w(S) = \sum_{i \in S} u_i$. A game v is the *minimum of a finite collection of games* (w^r) if for every coalition S, $v(S) = min_r w^r(S)$.

Kalai and Zemel (1982b) show that a game has nonempty subcores if and only if it is the minimum of a finite collection of additive games. Moreover, a game is such a minimum if and only if it is a flow game (as defined in section "Flow Games").

In summary, a game v in this important class of TU games can be characterized by any of the following five equivalent statements: (1) v has nonempty subcores, (2) v is totally balanced, (3) v is the minimum of additive games, (4) v is a market game, (5) v is a flow game.

Scarf (1967), Billera and Bixby (1973), and the follow-up literature extend some of the results above to general (NTU) games.

The Shapley TU Value

Definition 9 *The Shapley (1953) value of a TU game v is the payoff allocation $\varphi(v)$ defined by*

$$\phi_i(v) = \sum_{S:i \in S} \frac{(|S|-1)!(|N|-|S|)!}{N!}[v(S) - v(S/i)].$$

This expression describes the expected marginal contribution of player i to a random coalition. To elaborate, imagine the players arriving at the game in a random order. When player i arrives and joins the coalition of earlier arrivers S, he is paid his *marginal contribution* to that coalition, that is, $v(S \cup i) - v(S)$. His Shapley value $\phi_i(v)$ is the expected value of this marginal contribution when all orders of arrivals are equally likely.

Owen (1972) describes a parallel continuous-time process in which each player arrives at the game gradually. Owen extends the payoff function v to coalitions with 'fractionally present' players, and considers the instantaneous marginal contributions of each player i to such fractional coalitions. The Shapley value of player i is the integral of his instantaneous marginal contributions, when all the players arrive simultaneously at a constant rate over the same fixed time interval.

This continuous-time arrival model, when generalized to coalitional games with infinitely many players, leads to the definition of *Aumann–Shapley prices*. These are useful for the allocation of production costs to different goods produced in a nonseparable joint production process (see Tauman 1988; Young 1994).

A substantial literature is devoted to extensions and variations of the axioms that Shapley (1953) used to justify his value. These include extensions to infinitely many players and to general (NTU) games (discussed briefly below), and to nonsymmetric values (see Weber 1988; Kalai and

Samet 1987; Levy and McLean 1991;Monderer and Samet 2002, among others).

Is the Shapley value in the core of the game? Not always. But as Shapley (1971) shows, if the game is *convex*, meaning that $v(S \cup T) + v(S \cap T) \geq v(S) + v(T)$ for every pair of coalitions S and T, then the Shapley value and all the $n!$ profiles of marginal contributions (obtained under different orders of arrival) are in the core. Moreover, Ichiishi (1981) shows that the converse is also true.

We will turn to notions of value for NTU games after we describe solutions to the special case of two-person NTU games, that is, the Nash bargaining problem.

Solutions to Nash Bargaining Games

Nash (1950) pioneered the study of NTU games when he proposed a model of a two-person bargaining game and, using a small number of appealing principles, axiomatized the solution below.

Fix a two-person game V and for every imputation u define the *payoff gain* of player i by $gain_i(u) = u_i - v(i)$, with $v(i)$ being the highest payoff that player i can obtain on his own, that is, in his $V(i)$.

Definition 10 *The Nash bargaining solution is the unique imputation u that maximizes the product of the gains of the two players*, $gain_1(u) \cdot gain_2(u)$.

Twenty-five years later, Kalai and Smorodinsky (1975) and others showed that other appealing axioms lead to alternative solutions, like the two defined below.

The *ideal gain* of player i is $I_i = max_u gain_i(u)$, the maximum taken over all imputations u.

Definition 11 *The Kalai–Smorodinsky solution is the unique imputation u with payoff gains proportional to the players' ideal gains*, $gain_1(u)/gain_2(u) = I_1/I_2$.

Definition 12 *The egalitarian solution of Kalai* (1977a) *is the unique imputation u that equalizes the gains of the players,* $gain_1(u) = gain_2(u)$.

For additional solutions, including these of Raiffa (1953) and Maschler and Perles (1981), see the comprehensive surveys of Lensberg and Thomson (1989) and Thomson (1994).

Values of NTU Games

Three different extensions of the Shapley TU value have been proposed for NTU games: the *Shapley value* (extension), proposed by Shapley (1969) and axiomatized by Aumann (1985); the *Harsanyi value*, proposed by Harsanyi (1963) and axiomatized by Hart (1985); and the *egalitarian value*, proposed and axiomatized by Kalai and Samet (1985).

All three proposed extensions coincide with the original Shapley value on the class of TU games. For the class of NTU bargaining games, however, the (extended) Shapley value and the Harsanyi value coincide with the Nash bargaining solution, while the egalitarian value coincides with the egalitarian bargaining solution.

For additional material (beyond the brief discussion below) on these and related solutions, see McLean (2002).

Axiomatic Characterizations of Solutions

The imposition of general principles, or axioms, often leads to a unique determination of a solution. This approach is repeatedly used in game theory, as illustrated by the short summary below.

Nash's Axioms

Nash (1950) characterizes his bargaining solution by the following axioms: individual rationality, symmetry, Pareto optimality, invariance to utility scale, and independence of irrelevant alternatives (IIA).

Invariance to utility scale means that changing the scale of the utility of a player does not change the solution. But this axiom goes further by disallowing all methods that use information extraneous to the game, even if such methods are invariant to scale.

Nash's *IIA* axiom requires that a solution that remains feasible when other payoff profiles are removed from the feasible set should not be altered.

Shapley's Axioms

Shapley (1953) characterizes his TU value by the following axioms: symmetry, Pareto optimality, additivity, and dummy player.

A value is *additive* if in a game that is the sum of two games, the value of each player equals the sum of his values in the two component games.

A *dummy player*, that is, one who contributes nothing to any coalition, should be allocated no payoff.

Monotonicity Axioms

Monotonicity axioms describe notions of fairness and induce incentives to cooperate. The following are a few examples.

Kalai and Smorodinsky (1975) characterize their bargaining solution using *individual monotonicity*: a player's payoff should not be reduced if the set of imputations is expanded to improve his possible payoffs.

Kalai (1977a, b) and Kalai and Samet (1985) characterize their egalitarian solutions using *coalitional monotonicity*: expanding the feasible set of one coalition should not reduce the payoffs of any of its members.

Thomson (1983) uses *population monotonicity* to characterize the *n*-person Kalai-Smorodinsky solution: in dividing fixed resources among *n* players, no player should benefit if more players are added to share the same resources.

Maschler and Perles (1981) characterize their bargaining solution using *superadditivity* (used also in Myerson 1977a): if a bargaining problem is to be randomly drawn, all the players benefit by reaching agreement prior to knowing the realized game.

Young (1985) shows that Shapley's TU additivity axiom can be replaced by *strong monotonicity*: a player's payoff can only depend on his marginal contributions to his coalitions, and it has to be monotonically nondecreasing in these.

Axiomatizations of NTU Values

The NTU Shapley value is axiomatized in Aumann (1985) by adapting Shapley's TU axioms to the NTU setting, and combining them with Nash's IIA axiom. Different adaptations lead to an axiomatization of the Harsanyi (1963) value, as illustrated in Hart (1985). Kalai and Samet (1985) use coalitional monotonicity and a weak version of additivity to axiomatize the NTU egalitarian value.

For more information on axiomatizations of NTU values, see McLean (2002).

Consistency Axioms

Consistency axioms relate the solution of a game to the solutions of 'subgames' obtained when some of the players leave the game with their share of the payoff. Authors who employ consistency axioms include: Davis and Maschler (1965) for the bargaining set, Peleg (1985, 1986, 1992) for the core, Lensberg (1988) for the Nash *n*-person bargaining solution, Kalai and Samet (1987) and Levy and McLean (1991) for TU- and NTU-weighted Shapley values, Hart and Mas-Colell (1989) for the TU Shapley value, and Bhaskar and Kar (2004) for cost allocation in spanning trees.

Bridging Strategic and Coalitional Models

Several theoretical bridges connect strategic and coalitional models. Aumann (1961) offers two methods for reducing strategic games to coalitional games. Such reductions allow one to study specific strategic games, such as repeated games, from the perspectives of various coalitional solutions, such as the core.

One substantial area of research is the Nash program, designed to offer strategic foundations for various coalitional solution concepts. In Nash (1953), he began by constructing a strategic bargaining procedure, and showing that the strategic solution coincides with the coalitional Nash bargaining solution. See Nash program for a survey of the extensive literature that followed.

Network games and coalition formation are the subjects of a growing literature. Amending a TU game with a communication graph, Myerson (1977b) develops an appropriate extension of the Shapley value. Using this extended value,

Aumann and Myerson (1988) construct a dynamic strategic game of links formation that gives rise to stable communication graphs. For a survey of the large follow-up literature in this domain, see network formation.

Networks also offer a tool for the study of market structures. For example, Kalai et al. (1979) compare a market game with no restrictions to a star-shaped market, where all trade must flow through one middleman. Somewhat surprisingly, their comparisons of the cores of the corresponding games reveal the existence of economies in which becoming a middleman can only hurt a player.

Recent studies of strategic models of auctions point to interesting connections with the coalitional model. For example, empirical observations suggest that the better-performing auctions are the ones with outcomes in the core of the corresponding coalitional game. For related references, see Bikhchandani and Ostroy (2006), De Vries et al. (2007), and Day and Milgrom (2007).

Large Cooperative Games

When the number of players is large, the exponential number of possible coalitions makes the coalitional analysis difficult. On the other hand, in games with many players each individual has less influence and the laws of large numbers reduce uncertainties.

Unfortunately, the substantial fascinating literature on games with many players is too large to survey here, so the reader is referred to Aumann and Shapley (1974) and Neyman (2002) for the theory of the Shapley value of large games, and to Shapley and Shubik (1969a), Wooders and Zame (1984), Anderson (1992), Kannai (1992), and core convergence for the theory of cores of large games.

A surprising discovery drawn from the above literature is a phenomenon unique to large market games that has become known as the equivalence theorem: when applied to large market games, the predictions of almost all (with the notable exception of the von Neumann–Morgenstern stable sets) major solution concepts (in both coalitional and strategic game theory) coincide. Moreover, they all prescribe the economic price equilibrium as the solution for the game. This theorem presents the culmination of many papers, including Debreu and Scarf (1963), Aumann (1964), Shapley (1964), Shapley and Shubik (1969a) and Aumann (1975).

Directions for Future Work

Consider, for example, the task of constructing of a profit-sharing formula for a large consulting firm that has many partners with different expertise, located in offices around the world. While a coalitional approach should be suitable for the task, several current shortcomings limit its applicability. These include:

1. *Incomplete information*. Partners may have incomplete differential information about the feasible payoffs of different coalitions. While coalitional game theory has some literature on this subject (see Harsanyi and Selten 1972; Myerson 1984, and the follow-up literature), it is not nearly as developed as its strategic counterpart.
2. *Dynamics*. Although the feasible payoffs of coalitions vary with time, coalitional game theory is almost entirely static.
3. *Computation*. Even with a moderate number of players, the information needed for describing a game is very demanding. The literature on the complexity of computing solutions (as in Deng and Papadimitriou 1994; Nisan et al. 2007) is growing. But, overall, coalitional game theory is still far from offering readily computable solution concepts for complex problems like the profit-sharing formula in the situation described above.

Further research on the topics above would be an invaluable contribution to coalitional game theory.

See Also

- ▶ Cores
- ▶ Game Theory
- ▶ Matching and Market Design
- ▶ Shapley Value

Bibliography

The list below includes more than the relatively small number of papers discussed in this article, but due to space limitations many important contributions do not appear here.

Anderson, R.M. 1992. The core in perfectly competitive economies. In *The handbook of game theory with economic application*, ed. R.J. Aumann and S. Hart, Vol. 1. Amsterdam: North-Holland.

Aumann, R.J. 1961. The core of a cooperative game without side payments. *Transactions of the American Mathematical Society* 98: 539–552.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.

Aumann, R.J. 1975. Values of markets with a continuum of traders. *Econometrica* 43: 611–646.

Aumann, R.J. 1985. An axiomatization of the non-transferable utility value. *Econometrica* 53: 599–612.

Aumann, R.J., and S. Hart. 1992. *The handbook of game theory with economic application*. Vol. 1. Amsterdam: North-Holland.

Aumann, R.J., and S. Hart. 1994. *The handbook of game theory with economic application*. Vol. 2. Amsterdam: North-Holland.

Aumann, R.J., and S. Hart. 2002. *The handbook of game theory with economic application*. Vol. 3. Amsterdam: North-Holland.

Aumann, R.J., and M. Maschler. 1985. Game theoretic analysis of a bankruptcy problem from the Talmud. *Journal of Economic Theory* 36: 195–213.

Aumann, R.J., and M. Maschler. 1964. The bargaining set for cooperative games. In *Advances in game theory*, ed. M. Dresher, L.S. Shapley, and A.W. Tucker. Princeton: Princeton University Press.

Aumann, R.J., and R.B. Myerson. 1988. Endogenous formation of links between players and coalitions: An application of the Shapley value. In *The Shapley value*, ed. A. Roth. Cambridge: Cambridge University Press.

Aumann, R.J., and B. Peleg. 1960. Von Neumann–Morgenstern solutions to cooperative games without side payments. *Bulletin of the American Mathematical Society* 66: 173–179.

Aumann, R.J., and L.S. Shapley. 1974. *Values of non-atomic games*. Princeton: Princeton University Press.

Bhaskar, D., and A. Kar. 2004. Cost monotonicity, consistency and minimum cost spanning tree games. *Games and Econ Behavior* 48: 223–248.

Bikhchandani, S., and J.M. Ostroy. 2006. Ascending price Vickrey auctions. *Games and Econ Behavior* 55: 215–241.

Billera, L.J. 1970a. Existence of general bargaining sets for cooperative games without side payments. *Bulletin of the American Mathematical Society* 76: 375–379.

Billera, L.J. 1970b. Some theorems on the core of an n-person game without side payments. *SIAM Journal of Applied Mathematics* 18: 567–579.

Billera, L.J., and R. Bixby. 1973. A characterization of polyhedral market games. *International Journal of Game Theory* 2: 253–261.

Billera, L.J., D.C. Heath, and J. Raanan. 1978. Internal telephone billing rates: A novel application of non-atomic game theory. *Operations Research* 26: 956–965.

Binmore, K. 1987. Nash bargaining theory III. In *The economics of bargaining*, ed. K. Binmore and P. Dasgupta. Oxford: Blackwell.

Binmore, K., A. Rubinstein, and A. Wolinsky. 1986. The Nash bargaining solution in economic modelling. *Rand Journal of Economics* 17: 176–188.

Bird, C.G. 1976. On cost allocation for a spanning tree: A game theoretic approach. *Networks* 6: 335–360.

Bondareva, O.N. 1963. Some applications of linear programming methods to the theory of cooperative games. *Problemy Kibernetiki* 10: 119–139 [in Russian].

Brams, S.J., W.F. Lucas, and P.D. Straffin Jr. 1983. *Political and related models*. New York: Springer.

Chun, Y., and W. Thomson. 1990. Nash solution and uncertain disagreement points. *Games and Economic Behavior* 2: 213–223.

Davis, M., and M. Maschler. 1965. The kernel of a cooperative game. *Naval Research Logistics Quarterly* 12: 223–259.

Day, R., and P. Milgrom. 2007. Core selecting package auctions. *International Journal of Game Theory* 36: 393–407.

Debreu, G., and H. Scarf. 1963. A limit theorem on the core of an economy. *International Economic Review* 4: 236–246.

De Clippel, G., H. Peters, and H. Zank. 2004. Axiomatizing the Harsanyi solution, the symmetric egalitarian solution, and the consistent solution for NTU-games. *International Journal of Game Theory* 33: 145–158.

Deng, X., and C. Papadimitriou. 1994. On the complexity of cooperative game solution concepts. *Mathematics of Operations Research* 19: 257–266.

De Vries, S., J. Schummer, and R.V. Vohra. 2007. On ascending Vickrey auctions for heterogeneous objects. *Journal of Economic Theory* 132: 95–118.

Dubey, P., and L.S. Shapley. 1979. Some properties of the Banzhaf power index. *Mathematics of Operations Research* 4: 99–131.

Dubey, P., and L.S. Shapley. 1984. Totally balanced games arising from controlled programming problems. *Mathematical Programming* 29: 245–267.

Gale, D., and L.S. Shapley. 1962. College admissions and the stability of marriage. *American Mathematical Monthly* 69: 9–15.

Gillies, D.B. 1953. Some theorems on n-person games. Ph.D. thesis, Department of Mathematics, Princeton University.

Granot, D., and G. Huberman. 1984. On the core and nucleolus of the minimum costs spanning tree games. *Mathematical Programming* 29: 323–347.

Harsanyi, J.C. 1956. Approaches to the bargaining problem before and after the theory of games: A critical discussion of Zeuthen's, Hicks' and Nash's theories. *Econometrica* 24: 144–157.

Harsanyi, J.C. 1959. A bargaining model for the cooperative n-person game. In *Contributions to the theory of games*, ed. A.W. Tucker and R.W. Luce, Vol. 4. Princeton: Princeton University Press.

Harsanyi, J.C. 1963. A simplified bargaining model for the *n*-persons cooperative game. *International Economic Review* 4: 194–220.

Harsanyi, J.C. 1966. A general theory of rational behavior in game situations. *Econometrica* 34: 613–634.

Harsanyi, J.C., and R. Selten. 1972. A generalized Nash solution for two-person bargaining games with incomplete information. *Management Science* 18: 80–106.

Hart, S. 1973. Values of mixed games. *International Journal of Game Theory* 2: 69–86.

Hart, S. 1977a. Asymptotic values of games with a continuum of players. *Journal of Mathematical Economics* 4: 57–80.

Hart, S. 1977b. Values of non-differentiable markets with a continuum of traders. *Journal of Mathematical Economics* 4: 103–116.

Hart, S. 1980. Measure-based values of market games. *Mathematics of Operations Research* 5: 197–228.

Hart, S. 1985. An axiomatization of Harsanyi's nontransferable utility solution. *Econometrica* 53: 1295–1314.

Hart, S., and A. Mas-Colell. 1989. The potential: A new approach to the value in multiperson allocation problems. *Econometrica* 57: 589–614.

Ichiishi, T. 1981. Supermodularity: Applications to convex games and to the greedy algorithm for LP. *Journal of Economic Theory* 25: 283–286.

Jackson, M.O., and A. Wolinsky. 1996. A strategic model of social and economic networks. *Journal of Economic Theory* 71: 44–74.

Kalai, E. 1975. Excess functions for cooperative games without sidepayments. *SIAM Journal of Applied Mathematics* 29: 60–71.

Kalai, E. 1977a. Proportional solutions to bargaining situations: Interpersonal utility comparisons. *Econometrica* 45: 1623–1630.

Kalai, E. 1977b. Non-symmetric Nash solutions for replications of 2-person bargaining. *International Journal of Game Theory* 6: 129–133.

Kalai, E., A. Postlewaite, and J. Roberts. 1979. Barriers to trade and disadvantageous middlemen: Nonmonotonicity of the core. *Journal of Economic Theory* 19: 200–209.

Kalai, E., and R.W. Rosenthal. 1978. Arbitration of two-party disputes under ignorance. *International Journal of Game Theory* 7: 65–72.

Kalai, E., and D. Samet. 1985. Monotonic solutions to general cooperative games. *Econometrica* 53: 307–327.

Kalai, E., and D. Samet. 1987. On weighted Shapley values. *International Journal of Game Theory* 16: 205–222.

Kalai, E., and M. Smorodinsky. 1975. Other solutions to Nash's bargaining problems. *Econometrica* 43: 513–518.

Kalai, E., and E. Zemel. 1982a. Generalized network problems yielding totally balanced games. *Operations Research* 5: 998–1008.

Kalai, E., and E. Zemel. 1982b. Totally balanced games and games of flow. *Mathematics of Operations Research* 7: 476–478.

Kannai, Y. 1992. The core and balancedness. In *Handbook of game theory with economic applications*, ed. R.-J. Aumann and S. Hart, Vol. 1. Amsterdam: North-Holland.

Kaneko, M., and M. Wooders. 1982. Cores of partitioning games. *Mathematical Social Sciences* 3: 313–327.

Kohlberg, E. 1972. The nucleolus as a solution to a minimization problem. *SIAM Journal of Applied Mathematics* 23: 34–49.

Larauelle, A., and F. Valenciano. 2001. Shapley–Shubik and Banzhaf indices revisited. *Mathematics of Operations Research* 26: 89–104.

Lehrer, E. 1988. An axiomatization of the Banzhaf value. *International Journal of Game Theory* 17(2): 89–99.

Lensberg, T. 1985. Bargaining and fair allocation. In *Cost allocation, principles, applications*, ed. P. Young. Amsterdam: North-Holland.

Lensberg, T. 1988. Stability and the Nash solution. *Journal of Economic Theory* 45: 330–341.

Lensberg, T., and W. Thomson. 1989. *Axiomatic theory of bargaining with a variable population*. Cambridge: Cambridge University Press.

Levy, A., and R. McLean. 1991. An axiomatization of the non-symmetric NTU value. *International Journal of Game Theory* 19: 109–127.

Lucas, W.F. 1969. The proof that a game may not have a solution. *Transactions of the American Mathematical Society* 137: 219–229.

Lucas, W.F., and R.M. Thrall. 1963. n-person games in partition function forms. *Naval Research Logistics Quarterly* 10: 281–298.

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

Maschler, M. 1992. The bargaining set, kernel and nucleolus. In *Handbook of game theory with economic applications*, ed. R.J. Aumann and S. Hart, Vol. 1. Amsterdam: North-Holland.

Maschler, M., B. Peleg, and L.S. Shapley. 1979. Geometric properties of the kernel, nucleolus, and related solution concepts. *Mathematics of Operations Research* 4: 303–338.

Maschler, M., and M. Perles. 1981. The superadditive solution for the Nash bargaining game. *International Journal of Game Theory* 10: 163–193.

Mas-Colell, A. 1975. A further result on the representation of games by markets. *Journal of Economic Theory* 10: 117–122.

Mas-Colell, A. 1977. Competitive and value allocations of large exchange economies. *Journal of Economic Theory* 14: 419–438.

Mas-Colell, A. 1989. An equivalence theorem for a bargaining set. *Journal of Mathematical Economics* 18: 129–139.

McLean, R.P. 2002. Values of non-transferable utility games. In *Handbook of game theory with economic applications*, ed. R.J. Aumann and S. Hart, Vol. 3. Amsterdam: North-Holland.

Milnor, J.W., and L.S. Shapley. 1978. Values of large games II: Oceanic games. *Mathematics of Operations Research* 3: 290–307.

Monderer, D., and D. Samet. 2002. Variations on the Shapley value. In *Handbook of game theory with economic applications*, ed. R.J. Aumann and S. Hart, Vol. 3. Amsterdam: North-Holland.

Monderer, D., D. Samet, and L.S. Shapley. 1992. Weighted Shapley values and the core. *International Journal of Game Theory* 21: 27–39.

Moulin, H. 1988. *Axioms of cooperative decision making*. Cambridge: Cambridge University Press.

Myerson, R.B. 1977a. Two-person bargaining problems and comparable utility. *Econometrica* 45: 1631–1637.

Myerson, R.B. 1977b. Graphs and cooperation in games. *Mathematics of Operations Research* 2: 225–229.

Myerson, R.B. 1979. Incentive compatibility and the bargaining problem. *Econometrica* 47: 61–74.

Myerson, R.B. 1984. Cooperative games with incomplete information. *International Journal of Game Theory* 13: 69–96.

Myerson, R.B. 1991. *Game theory: Analysis of conflict*. Cambridge: Harvard University Press.

Nash, J.F. 1950. The bargaining problem. *Econometrica* 18: 155–162.

Nash, J.F. 1953. Two person cooperative games. *Econometrica* 21: 128–140.

Neyman, A. 1985. Semivalues of political economic games. *Mathematics of Operations Research* 10: 390–402.

Neyman, A. 1987. Weighted majority games have an asymptotic value. *Mathematics of Operations Research* 13: 556–580.

Neyman, A. 2002. Values of games with infinitely many players. In *Handbook of game theory with economic applications*, ed. R.J. Aumann and S. Hart, Vol. 3. Amsterdam: North-Holland.

Nisan, N., T. Roughgarden, E. Tardos, and V. Vazirani. 2007. *Algorithmic game theory.* Cambridge: Cambridge University Press.

O'Neil, B. 1982. A problem of rights arbitration from the Talmud. *Mathematical Social Sciences* 2: 345–371.

Owen, G. 1972. Multilinear extensions of games. *Management Science* 18: 64–79.

Owen, G. 1975. On the core of linear production games. *Mathematical Programming* 9: 358–370.

Osborne, M.J., and A. Rubinstein. 1994. *A course in game theory.* Cambridge, MA: MIT Press.

Peleg, B. 1963a. Solutions to cooperative games without side payments. *Transactions of the American Mathematical Society* 106: 280–292.

Peleg, B. 1963b. Bargaining sets of cooperative games without side payments. *Israel Journal of Mathematics* 1: 197–200.

Peleg, B. 1985. An axiomatization of the core of cooperative games without side payments. *Journal of Mathematical Economics* 14: 203–214.

Peleg, B. 1986. On the reduced games property and its converse. *International Journal of Game Theory* 15: 187–200.

Peleg, B. 1992. Axiomatizations of the core. In *Handbook of game theory with economic applications*, ed. R.-J. Aumann and S. Hart, Vol. 1. Amsterdam: North-Holland.

Peleg, B., and P. Sudholter. 2003. *Introduction to the theory of cooperative games*. Dordrecht: Kluwer Academic Publications.

Peters, H.J.M. 1992. *Axiomatic bargaining game theory.* Dordrecht: Kluwer Academic Publishers.

Peters, H., S. Tijs, and A. Zarzuelo. 1994. A reduced game property for the Kalai Smorodinsky and Egalitarian bargaining solution. *Mathematical Social Sciences* 27: 11–18.

Potters, J., I. Curiel, and S. Tijs. 1992. Traveling salesman game. *Mathematical Programming* 53: 199–211.

Raiffa, H. 1953. Arbitration schemes for generalized two-person games. In *Contributions to the theory of games II*, ed. H. Kuhn and A.W. Tucker. Princeton: Princeton University Press.

Riker, W.H., and L.S. Shapley. 1968. Weighted voting: A mathematical analysis for instrumental judgements. In *Representation*, ed. J.R. Pennock and J.W. Chapman. New York: Atherton.

Roth, A.E. 1977. The Shapley value as a von Neumann–Morgenstern utility. *Econometrica* 45: 657–664.

Roth, A.E. 1979. *Axiomatic models of bargaining*. Berlin/New York: Springer.

Roth, A.E. 1984. The evolution of the labor market for medical interns and residents: A case study in game theory. *Journal of Political Economics* 92: 991–1016.

Roth, A.E., and R.E. Verrecchia. 1979. The Shapley value as applied to cost allocation: A reinpt. *Journal of Accounting Research* 17: 295–303.

Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50: 97–109.

Scarf, H.E. 1967. The core of an n-person game. *Econometrica* 35: 50–69.

Schmeidler, D. 1969. The nucleolus of a characteristic function game. *SIAM Journal of Applied Mathematics* 17: 1163–1170.
Schmeidler, D. 1972. Cores of exact games I. *Journal of Mathematical Analysis and Application* 40: 214–225.
Shapley, L.S. 1952. *Notes on the n-person game iii: Some variants of the Von Neumann–Morgenstern definition of solution*. Vol. 817. Santa Monica: RAND Corporation RM.
Shapley, L.S. 1953. A value for n-person games. In *Contributions to the theory of games II*, ed. H. Kuhn and A.W. Tucker. Princeton: Princeton University Press.
Shapley, L.S. 1964. *Values of large games, VII: A general exchange economy with money*. Vol. 4248. Santa Monica: RAND Corporation RM.
Shapley, L.S. 1967. On balanced sets and cores. *Naval Research Logistics Quarterly* 14: 453–460.
Shapley, L.S. 1969. Utility comparison and the theory of games. In *La Décision*. Paris: Edition du CNRS.
Shapley, L.S. 1971. Cores of convex games. *International Journal of Game Theory* 1: 11–26.
Shapley, L.S. 1973. Let's block 'block'. *Econometrica* 41: 1201–1202.
Shapley, L.S., and M. Shubik. 1954. A method for evaluating the distribution of power in a committee system. *American Political Science Review* 48: 787–792.
Shapley, L.S., and M. Shubik. 1969a. On market games. *Journal. of Economic Theory* 1: 9–25.
Shapley, L.S., and M. Shubik. 1969b. Pure competition, coalitional power and fair division. *International Economic Review* 10: 337–362.
Shubik, M. 1959. *Strategy and market structure: Competition, oligopoly, and the theory of games*. New York: Wiley.
Shubik, M. 1962. Incentives, decentralized control, the assignment of joint costs and internal pricing. *Management Science* 8: 325–343.
Shubik, M. 1982. *Game theory in the social sciences: Concepts and solutions*. Cambridge, MA: MIT Press.
Shubik, M. 1984. *A game theoretic approach to political economy*. Cambridge, MA: MIT Press.
Sprumont, Y. 1998. Ordinal cost sharing. *Journal of Economic Theory* 81: 26–162.
Tauman, Y. 1981. Value on a class of non-differentiable market games. *International Journal of Game Theory* 10: 155–162.
Tauman, Y. 1988. The Aumann-Shapley prices: A survey. In *The Shapley value*, ed. A. Roth. New York: Cambridge University Press.
Thomson, W. 1983. The fair division of a fixed supply among a growing population. *Mathematics of Operations Research* 8: 319–326.
Thomson, W. 1987. Monotonicity of bargaining solutions with respect to the disagreement point. *Journal of Economic Theory* 42: 50–58.
Thomson, W. 1994. Cooperative models of bargaining. In *Handbook of game theory with economic applications*, ed. R.J. Aumann and S. Hart, Vol. 3, 1237–1284. Amsterdam: North-Holland.
Valenciano, F., and J.M. Zarzuelo. 1994. On the interpretation of the nonsymmetric bargaining solutions and their extensions to nonexpected utility preferences. *Games and Economic Behavior* 7: 461–472.
Von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.
Weber, R.J. 1988. Probabilistic values for games. In *The Shapley value*, ed. A. Roth. Cambridge: Cambridge University Press.
Weber, R.J. 1994. Games in coalitional form. In *Handbook of game theory with economic applications*, ed. R.-J. Aumann and S. Hart, Vol. 2. Amsterdam: North-Holland.
Wilson, R. 1978. Information, efficiency, and the core of an economy. *Econometrica* 46: 807–816.
Winter, E. 2002. The Shapley value. In *Handbook of game theory with economic applications*, ed. R.J. Aumann and S. Hart, Vol. 3. Amsterdam: North-Holland.
Wooders, M.H., and W.R. Zame. 1984. Approximate cores of large games. *Econometrica* 52: 1327–1350.
Wooders, M.H., and W.R. Zame. 1987. Large games: Fair and stable outcomes. *Journal of Economic Theory* 42: 59–63.
Young, H.P. 1985. Monotonic solutions of cooperative games. *International Journal of Game Theory* 14: 65–72.
Young, H.P. 1994. Cost allocation. In *Handbook of game theory with economic applications*, ed. R.J. Aumann and S. Hart, Vol. 2. Amsterdam: North-Holland.

Games with Incomplete Information

Robert J. Weber

Classical economic models almost universally assume that the resources and preferences of individuals (or firms) are known not only to the individuals themselves but also to their competitors. In practice, this assumption is rarely correct. Once the attempt is made to include uncertainty (not just about the environment but also about other strategic actors) within economic models, it becomes necessary to broaden those models substantially, to include considerations about the beliefs of individuals concerning the status of their competitors, as well as about learning as it takes place over time. A standard approach for doing this is to model the situation under investigation as a game with incomplete

information, and to study the (Bayesian) equilibrium points of that game.

This approach has been used in recent years to analyse such issues as negotiation, competitive bidding, social choice, limit pricing, the signalling roles of education and advertising, together with a variety of other phenomena which arise under the general heading of industrial organization.

Games in Strategic Form

Consider first games in strategic form, wherein the competitors each must choose a single action. In principle, any game can be reduced to this form by letting the actions available to the players be sufficiently complex (e.g. poker can be modelled in this manner).

An *n*-player *game with incomplete information* consists of the following elements: (1) for each player *i*, a probability space *Ti* of that player's possible types, a set *Ai* of actions available to that player, and a pay-off function *ui* defined for every combination $(t, a) = (t1, \ldots, tn, an)$ of player types and actions; and (2) a probability measure μ on the space. It is assumed that the elements of the game are commonly known to the players. At the start of the game, the *n*-tuple of player types is determined according to ü. Each player is privately informed of his own type, and then the players simultaneously announce their chosen actions. Each player finally receives the pay-off corresponding to the combination (t, a) of types and announced actions.

For example, assume that each player in a game knows his own preferences but is uncertain about the preferences (and hence, about the strategic motivations) of his competitors. This situation may be modelled as a game in which the pay-off functions have the form $ui(ti, a)$. The realization of the random variable *t*i is player *i*'s type, known to him but unknown to the other players.

In contrast, assume that the preferences of the players are known to all, but that the payoffs are affected by some chance event represented by the random variable *to*; that is, each payoff function can be written in the form $v_i(t_o, a)$. The variable *ti* represents a private signal received by player *i* prior to his choice of an action. Note that a player's signal may be informative about the signals of the others, as well as just about the chance event, through the joint distribution of $(t_o \langle t_1, \ldots, t_n \rangle$. In this case, the expected pay-off of a player, given that the vector $t = (t_1, \ldots, t_n)$ of signals has arisen and the players have selected the actions $a = (a_1, \ldots, a_n)$, is $u_i(t, a) = E[v_i(t_o, a)|t_1, \ldots, t_n]$.

The Notion of 'Type'

The type-based formulation of a game with incomplete information is due to Harsanyi (1967–8), who proposed it as a way of cutting through the complexities of modelling not only a player's information and preferences but also his beliefs about other players' information and preferences, and his beliefs about their beliefs, and so on.

Mertens and Zamir (1985) subsequently presented a formulation of games with incomplete information which unifies the type-based approach with the beliefs-about-beliefs (and so on) approach to settings of incomplete information. By specifically modelling the interated sequence of beliefs which determines a player's state of knowledge at the beginning of the game, and then considering 'consistent beliefs-closed subspaces' of the general space of players' beliefs, they were able to show that the original Harsanyi formulation involves no essential loss of generality.

Strategies and Equilibria

A *strategy* for a player specifies the action (or randomized choice of action) to be taken by each potential type of that player. The action specificed for his actual type can be thought of as his 'private strategy'. In practice, even when a player has already learned his type, in order to decide upon his own appropriate action he must form a hypothesis concerning the strategies to be used by the others. But to analyse their strategic

problems, he must ask himself what strategy they will expect him to follow. Therefore, it is necessary for him to consider the strategic choices his other potential types would make, in order to select an appropriate action for his actual type.

A (Bayesian) *equilibrium point* of a game is an n-tuple of strategies, in which the private strategy of each type of each player is a best response for that type of the $(n-1)$-tuple of strategies specified for the other players. This definition directly generalizes that of a Nash equilibrium point for a game with complete information.

As an example, consider two individuals who jointly own a piece of land. They have decided to sever their relationship, and for one of the two to buy the land from the other. Each knows how valuable the land is to himself, but is unsure of its worth to the other. They agree that each will write down a bid; the high bidder will keep the land and will pay the amount of his bid to the other.

Assume that each is equally likely to value the land at any level between \$0 and \$1200, and that both know this. At the unique Bayesian equilibrium point of the bidding game, each bids one-third of his own valuation. If, for example, one of them values the land at \$300 and believes the other to be following the indicated equilibrium strategy, then by bidding \$100 he has an expected pay-off of $1/4 \cdot \$200 + 3/4 \cdot \$250$; that is, he expects to win with probability 1/4, and when he loses, he expects the other's (winning) bid to be between \$100 and \$400. This private strategy is optimal for him, given his belief about the other's behaviour. More generally, given his belief that his partner will bid a third of the partner's valuation, his own expected pay-off, when his valuation is v and he bids b, is $(3b/1200).(v-b) + (1-3b/1200).(b+400)/2$. This is maximized by taking $b = v/3$.

Distributional Strategies

In order to study the sensitivity of equilibrium results to variations in the informational structure of a game, it is necessary to define topologies on both the spaces of player strategies and the space of games. The first may be done by recasting the definition of a strategy in distributional form:

A *distributional strategy* v for a player is a probability measure on the product of his type and action spaces, with the property that the marginal distribution of v on the player's type space coincides with the original marginal distribution induced by μ. Player i, knowing his type t_i, chooses his action according to the conditional distribution $v(\cdot|t_i)$; an outside observer, seeing the player's action a_i, will revise his beliefs concerning the players type to $v(\cdot|a_i)$. A natural topology on a player's strategy space v is the topology of weak convergence of probability measures.

Taking this distributional perspective, Milgrom and Weber (1985) proved a general equilibrium existence theorem; in particular, it follows from this theorem that any game with compact action spaces, uniformly continuous pay-off functions, and for which the type distribution is absolutely continuous with respect to the product of the marginal type distributions (i.e. for which the joint distribution of types has a corresponding joint probability density function), has an equilibrium point in distributional strategies. They also showed that, with the appropriate topology defined on the space of games, any limit point of equilibria of a sequence of games is an equilibrium point of the limit game. One consequence of the distributional approach is that when the games in a sequence provide a player with private information which disappears in the limit game, a sequence of pure strategies for that player can converge to a randomized strategy in the limit game. This reinforces an observation first offered by Harsanyi (1973) to explain why, in practice, decision-makers are rarely observed to randomize their choices of actions: the existence of a slight amount of private information is sufficient, in most cases, to allow the decision-makers to follow pure strategies which present, to their competitors, the appearance of a randomized choice of actions. In essence, competitors observe the marginal distribution, induced by a player's distributional strategy, on his action space.

Inefficiencies Created by Incentive Constraints

In many circumstances, parties holding private information can find it difficult, or even impossible, to arrange efficient trades. A simple example, drawn from a class of problems first discussed by Akerlof (1970), concerns the owner of a car, attempting to arrange the sale of that car to a prospective buyer. Assume that the value of the car to the seller is primarily based on the quality of the car, and that the seller knows this value. Further assume that, whatever the car is worth to the seller, it is worth 50 per cent more to the buyer. And finally, assume that the buyer's only knowledge about the seller's value is that it is uniformly distributed between \$0 and \$1000. In this case, it is commonly known to the two parties that a mutually advantageous trade exists. Nevertheless, no sale can be expected to take place, since the seller's willingness to accept any price \$$x$ signals to the buyer that the seller's valuation lies between \$0 and \$$x$, and therefore that the expected value of the car to the buyer is most likely no more than $3/2 \cdot (x/2) = 3/4 \cdot x$. As long as the initial uncertainty persists (i.e. as long as no pre-sale verification of the car's quality is possible), and as long as no contingent trade can be arranged (i.e. as long as no warranty can be written), trade is impossible – even if the parties agree to consult an intervenor.

Intervenors in settings of incomplete information typically act as game designers, influencing the flow of information between parties, enforcing agreements and in some cases actually specifying the final resolution of a dispute (e.g. binding arbitration). Essentially, an intervenor creates a game which the parties must play. Any theory of intervention must therefore be tied to the issue of designing games with desirable equilibrium outcomes.

The Akerlof example shows that if intervenors are restricted from playing an auditing role, and if the outcome of the game cannot be made contingent on the parties' true types, then ex-post inefficient outcomes are at times inevitable. This understanding has led to the development of the theory of 'incentive-efficient mechanism design'.

The Revelation Principle

In the area of game design, a simple, yet conceptually deep, type of analysis has become standard. Consider any equilibrium pair of strategies in a particular two-person game. (The following analysis is equally valid for games involving more than two players.) Each party's strategy can be viewed as a book, with each chapter detailing the private strategy of one that party's types. Given the two actual types, a pairing of the private strategies in the appropriate chapters of the two books will lead to an outcome of the game.

Now, step back from this setting and imagine the two parties in separate rooms, each instructing an agent on how to act on his behalf. Each agent holds in hand the strategy book of his side; all he must be told is which chapter to use. From this new perspective, the original two parties can be thought of as playing an 'agent-instruction' game, in which the strategy books are prespecified and each must merely tell his agent his type (or, equivalently, point to a chapter in his strategy book). An equilibrium point in this new 'type-revelation' game is for each to tell the truth to his agent. Otherwise, the original strategies could not have been in equilibrium in the original game. Consequently, anything which can be accomplished at equilibrium through the use of any particular dispute-resolution procedure can also be accomplished through the use of some other procedure in which the only actions available to the parties are to state their (respective) types, and in which it is in equilibrium for each to reveal his type truthfully.

This observation, known as the 'revelation principle', reduces the problem of game design to the problem of optimizing the designer's objective function, subject to a collection of 'incentive constraints', one for each type of each player. An early application of this approach was to the design of auction procedures which maximize the seller's expected revenue. Myerson (1984) subsequently applied the approach to the problem of bargaining under uncertainty, and provided a generalization of the classical complete-information Nash bargaining solution. A central feature of this generalization is the incorporation of intrapersonal (i.e. intertype) equity considerations.

Games in Extensive Form

A game with incomplete information in extensive form begins with a chance move which determines the types of the players, and continues with an information structure which preserves the privacy of each player's information. Many multi-stage bargaining problems can be represented in this form; typically, such games have a large number of equilibria, including equilibria in which one party is completely intransigent and the other concedes immediately, as well as equilibria in which both parties make information-revealing concessions over the series of stages.

A classical approach to the identification of 'plausible' equilibria in games with complete information is to seek equilibria which are subgame-perfect; that is, which specify optimal actions for all parties in all subgames of the original game. For example, Rubinstein (1982) presented a repeated offer-counteroffer game with many equilibria, and demonstrated that the requirement of subgame perfection uniquely identified one of those equilibria. However, subgame perfection is a concept of little use in distinguishing between equilibria of a game with incomplete information, since the privacy of the players' information typically results in the original game having no proper subgames.

Selten (1975), with his notion of 'trembling-hand' perfection, and Kreps and Wilson (1982), with their closely related notion of sequential equilibrium, provided extensions of the concept of subgame perfection which require that players act optimally at positions off the equilibrium path of the game. Central to the Kreps–Wilson approach is the incorporation of players' interim beliefs (about the other players' types, and past and future actions) at all game positions in the specification of an equilibrium point. Subsequent work on equilibrium selection in games with incomplete information has relied heavily on the study of justifiable out-of-equilibrium beliefs.

Repeated Games

A special kind of extensive-form game consists of an initial chance move which determines the players' types, followed by the repeated play of a single game with type-dependent pay-offs. Players are not allowed to observe the actual stage-to-stage pay-offs during play, but are allowed to monitor the stage-to-stage actions of their competitors. The study of such games provides insight into the way players learn about one another over time; that is, insights into the way reputations are developed and maintained or changed.

Beginning in 1965 with research sponsored by the US Arms Control and Disarmament Agency, substantial effort has been focused on the study of infinitely repeated games with incomplete information. A principal result in the two-person, zero-sum case is that optimal strategies typically involve a single initial reference to the information a party holds, followed by period-to-period moves which depend only on the outcome of that single reference. (In an infinitely repeated game, short-term pay-offs are unimportant. Whatever behaviour a player adopts, his opponent's beliefs will converge to some limit; the long-term pay-offs will depend only on the limiting beliefs of the players. Therefore, in a strictly competitive environment it is sufficient for a player to determine at the beginning of the game precisely how much information he will eventually reveal.) Hart (1985) extended this analysis to games with private information on one side, and gains available to the players through cooperative actions. His work demonstrates that, when mutual gains are available, equilibrium behaviour may involve a series of references by the informed player to his information, interspersed with joint randomizing actions between the players which determine what information will next be revealed.

For many years, the finitely repeated Prisoners' Dilemma posed a dilemma for game theorists. Set in the framework of complete information, this game has a unique equilibrium outcome: the players never cooperate with one another. However, experiments repeatedly showed that actual players frequently establish a pattern of cooperation which persists until the game approaches its final stage. Kreps et al. (1982) finally offered an explanation for this discrepancy, by demonstrating that a slight

change in the initial informational framework yields games with equilibrium outcomes similar to the observed experimental outcomes. For example, assume that each player initially assigns a small positive probability to his opponent being the type of individual who (irrationally) will always respond to cooperation in one stage with further cooperation in the next. Then there will be equilibria in which, even when both players are actually rational, they will (with high probability) cooperate until near the end of the game. An interpretation of such equilibrium behaviour is that each finds it to his benefit to build a reputation as the irrational, cooperative type. The incomplete information model is necessary to obtain this behaviour. If the initial uncertainty as to type did not exist in the mind of a player's opponent, such a reputation would be impossible to build. An emerging 'theory of reputation' has its roots in this analysis.

See Also

- ▶ Game Theory
- ▶ Incomplete Contracts
- ▶ Oligopoly and Game Theory

Bibliography

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

Harsanyi, J.C. 1967–8. Games with incomplete information played by Bayesian players. *Management Science* 14: 159–82, 320–34, 486–502.

Harsanyi, J.C. 1973. Games with randomly-distributed payoffs: A new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory* 2: 1–23.

Hart, S. 1985. Nonzero-sum two-person repeated games with incomplete information. *Mathematics of Operations Research* 10: 117–153.

Kreps, D.M., and R. Wilson. 1982. Sequential equilibria. *Econometrica* 50: 863–894.

Kreps, D.M., P. Milgrom, J. Roberts, and R. Wilson. 1982. Rational cooperation in the finitely repeated Prisoner's dilemma. *Journal of Economic Theory* 27: 245–252.

Mertens, J.-F., and S. Zamir. 1985. Formulation of Bayesian analysis for games with incomplete information. *International Journal of Game Theory* 14: 1–29.

Milgrom, P.R., and R.J. Weber. 1985. Distributional strategies for games with incomplete information. *Mathematics of Operations Research* 10: 619–632.

Myerson, R.B. 1984. Two-person bargaining problems with incomplete information. *Econometrica* 52: 461–487.

Rubinstein, A. 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50: 97–109.

Selten, R. 1975. Reexamination of the perfectness concept for equilibriu points in extensive games. *International Journal of Game Theory* 4: 25–55.

Gaming Contracts

E. Schuster

> A gaming or wagering contract is one by which two persons professing to hold opposite views touching a future uncertain event mutually agree that, dependent upon that event, one shall receive from the other, and the other shall pay or hand over to him, a sum of money or other stake; neither of the contracting parties having any other interest in that contract than the sum or stake he will so win or lose, there being no other real consideration for the making of such contract by either of the parties (Justice Hawkins in *Carlill* v. *Carbolic Smoke Ball Company* (1892), Queen's Bench 484).

Contracts of this description are declared to be null and void by an act passed during the present reign [of Queen Victoria] (8 & 9 Vict. c. 109. p. 18). Notwithstanding this act it was held that a betting agent who had paid the amount due by his principal on the loss of a bet was entitled to recover the same from the latter (*Read* v. *Anderson*, 10 Queen's Bench Division 100; 13 Queen's Bench Division 779), but this indirect recognition of betting transactions has lately been set aside by the Gaming Act of 1892, which enacts that no action shall be brought to recover any sum of money paid in respect of any gaming or wagering contract. The subject of gaming contracts has recently been discussed with reference to the 'missing word competitions' organized by certain newspapers, which were held to be illegal, as the result did not depend on skill and judgement but upon mere chance (*Barclay* v. *Pearson* (1893),

2 Chancery 154). The principle of the statute against gaming and wagering contracts was, to a certain extent, already recognized by the statute of 14 Geo. III. c. 48, which forbids the insurance of a life in which the insurer has no interest, and which is still in force.

Much discussion has taken place both in England and abroad on the question whether certain time bargains on the stock exchange and in the produce markets are to be considered as partaking of the nature of wagers, and the result of the decisions seems to be that a contract is not enforceable where it can be proved that it was not the intention of the parties to deliver or receive a certain quantity of securities or produce at a certain price, but that the payment of the difference between the price at which the bargain was made and the market price at the time fixed for the completion of the bargain was the sole object of the transaction; in the absence of such proof the parties must be presumed to have intended a real sale.

Ganilh, Charles (1785–1836)

P. Bridel

Ganilh was born in Allanches (Cantal) on 6 January 1785, and died in Paris in 1836. The main claims to (modest) fame of this barrister-turned-politician-turned-economist are his *Systèmes d'économie politique* (1809) and his *Dictionnaire d'économie politique* (1826). These two works are respectively the first systematic history of thought and the first theoretically oriented dictionary of economics ever published. Unfortunately, save this claim for priority, these two prestigious titles barely conceal the analytical poverty of their contents.

Ganilh's two main analytical books can best be seen as potted (though not uncritical) introductions of the main Smithian theses to the French educated layman. In the *Systèmes*, leaving aside all anecdotal reference to economists and their intellectual environment, Ganilh concentrates exclusively on a history of economic theory. Centred around main concepts (wealth, labour, value, capital, money etc.), Ganilh's systematic treatment ranges from the Greek philosophers down the centuries to the Mercantilist and Physiocratic schools which, together with the *Wealth of Nations*, are given the pride of place. In a similar way, in the *Dictionnaire*, what Ganilh considers to be the main theoretical concepts (excluding any factual or biographical entries) are not only individually discussed in alphabetic order but also logically connected by means of a cross-reference system.

Free trade and the notion of productive labour are the two themes around which Ganilh's resistance to Adam Smith are articulated. Without reverting to the Mercantilist doctrine that dominated French regulatory practice right up to the fall of the *ancien régime*, Ganilh thinks fit (for practical reasons) to bring some restriction to the then newly discovered free trade doctrine. Anticipating in some ways Frederic List's *National System of Political Economy* (1841), Ganilh strongly advocates trade barriers in favour of France's nascent industries (1826, pp. 142–150). For his critical remarks on Smith's concept of *productive labour* (i.e. labour supported by capital to produce material goods), Ganilh was caught in a heated argument with *inter alia* Malthus, Ricardo, Buchanan and Lauderdale (1826, pp. 102–4 and 413–28). With his opinion that 'any labour the exchange of which gives rise to a value' (1826, p. 415) is productive, Ganilh espoused the then dominant utilitarian doctrine of Say; and for that, he was later to be disparagingly condemned by Marx.

In addition to these two analytical works, Ganilh wrote extensively in the field of public finance (1815).

Selected Works

1809. *Des systèmes d'économie politique*, 2 vols. Paris: Déterville.
1815. *Théorie de l'économie politique*, 2 vols. Paris: Déterville.
1826. *Dictionnaire analytique de l'économie politique*. Paris: Ladvocat.

Garnier, Clément Joseph (1813–1881)

R. F. Hébert

French economist, born at Beuil (Alpes Maritimes) on 3 October 1813; died at Paris on 25 September 1881. Joseph Garnier (not to be confused with the translator of Adam Smith) came from a family of prosperous farmers, but showed no inclination to follow his ancestral heritage. He made his way to Paris in 1829, when only 16. Poised to join the banking firm of Lafitte, he was induced instead to enter the Ecole Supérieure de Commerce by a family friend, Adolphe Blanqui. Later following in Blanqui's footsteps. Garnier became both teacher and principal at the Ecole.

Garnier remained in the mainstream of French economics throughout the mid–19th century. He was one of the founders of the Société d'Economie Politique and its permanent secretary from 1842 until his death in 1881. A tireless teacher, he held professorships at five different schools. In 1843 he began a series of lectures at the Athenée, which eventually evolved into his *Eléments de l'économie politique*, a popular and encyclopedic treatise that went through multiple editions from 1846 to 1907. The work was retitled in its fourth edition to the *Traité de l'économie politique*, and in this form it was eventually translated into Spanish, Italian and Russian. Already on the strength of the first edition, Garnier had been named in 1846 to the newly created chair of political economy at the École des Ponts et Chaussées (Dupuit's *alma mater*), a position he held for 35 years. He also spent a quarter of a century as the chief editor of the *Journal des économistes*, the leading French economics journal of the period.

His fortuitous placement at the nerve centres of French economics (the faculty of a *grande école*, the leading journal and professional society) gave Garnier a measure of influence beyond what could be expected from the originality of his ideas, which in any case was minimal. He was chiefly an exponent of classical economics, occupying the middle ground between the virulent strain of liberalism that afflicted Molinari and the less strident version associated with Blanqui and Dunoyer. Although he produced a popular, annotated French edition of Malthus's *Essay on Population*, Garnier chose in his own works to expand the optimistic doctrines of Smith and Say rather than the underside of Ricardo and Malthus. His *Traité* is a good example of French economics before Mill. Its popularity must be attributed in part to this orthodoxy, but also to Garnier's depth of knowledge and his orderly presentation.

Garnier was named a member of the French Institute in 1873. Three years later he was elected to the French Senate by voters in his home district, thus capping a long academic and literary career with a final dimension of public service.

Selected Works

1846. *Eléments de I'économie politique, exposé des notions fondamentales de cette science*. Paris: Guillaumin.

1858a. *Premiéres notions d'économie politique, sociale ou industrielle*. Paris: Guillaumin.

1858b. *Eléments de finances suivis de éléments de statistique, de la misére, l'association et l'économie politique*. Paris: Garnier frères.

Gasoline Markets

Jean-François Houde

Abstract

Gasoline retail markets have traditionally attracted a lot of attention from researchers and policy makers. This entry reviews a set of questions studied by Industrial Organisation (IO) economists related to this market. The discussion is organised around three themes. The first reviews papers concerned

with the transmission of cost shocks and the cyclical properties of gasoline prices. The second theme includes a variety of papers testing for market power, and providing evidence in favour of price discrimination, tacit collusion, differentiation and vertical restraints. Finally, the last section is devoted to papers evaluating the consequences of economic and environmental regulations in gasoline markets.

Keywords

Gasoline; Retailing; Oligopoly; Pricing; Regulation; Vertical relations

Within the petroleum industry, gasoline markets have been a focus of intensive research. The availability of large datasets, combined with a complex and diverse industrial structure, has motivated many important empirical and theoretical articles. This entry will focus on the retail segment of the petroleum industry, which also includes extraction, refinery and distribution.

Despite often being portrayed as the archetype of a perfectly competitive market, economists and policymakers have long been intrigued by the behaviour of gasoline prices. It is not uncommon to observe prices jumping sharply in a coordinated manner, or crude oil price shocks being passed through asymmetrically to consumers. Gasoline price wars have attracted a great deal of attention as well, sometimes triggering price control regulations from governments in the USA and around the world. The recent wave of vertical mergers has led the US senate to conduct a comprehensive analysis of the industry, as summarised by Senator Carl Levin's report in 2002. Consumers are also greatly concerned about gasoline price movements, as it occupies a between 4.5% and 12.4% of households' disposable income, especially for poorer households (Gicheva et al. 2007).

This entry will attempt to summarise the academic research studying gasoline markets. The discussion is organised around three themes: (i) incomplete passthrough and retail price cycles, (ii) evaluation of market power, and (iii) regulation.

Incomplete Pass-Through and Retail Price Cycles

Like many commodity goods, gasoline prices are known to adjust slowly to cost changes. What is more intriguing is the fact that gasoline prices commonly adjust faster to cost increases than cost decreases. This phenomenon, known as *rockets and feathers*, was first documented by Bacon (1991) in an analysis of the UK gasoline market. Peltzman (2000) documents similar asymmetries in a wide array of retail markets.

Among the most detailed studies, Borenstein et al. (1997) demonstrate the existence of a large asymmetric pass-through in US markets: retail prices are on average 0.55 cents *higher* two weeks after a 1 cent increase in the price of crude oil, but are only 0.18 cents *lower* after a 1 cent decrease. They also show that more than half of this asymmetry is due the adjustment of wholesale prices, while the remainder is due to a small but significant asymmetric adjustment of retail prices.

The most likely explanation for the upstream asymmetry is related to the interplay of production adjustment costs and storage capacity at the refinery level (see Borenstein and Shepard (2002)). At the retail level, several theories have been proposed. It is clear for instance that asymmetric adjustments are more prevalent in concentrated retail markets (Deltas 2008), and Borenstein et al. have argued that tacit collusion can cause the retail price asymmetries.

Information frictions between consumers and firms have also been proposed to explain the phenomenon. Tappata (2009), for instance, builds an oligopoly pricing model with search frictions in which consumers endogenously choose a lower search effort when costs are expected to be high. This leads to a relatively inelastic demand after an unexpected decline in wholesale prices. Yang and Ye (2008) and Lewis (2005) also provide related search-based theoretical and empirical analysis of asymmetric pass-through.

A related phenomenon has gained a lot of attention. In many cities gasoline prices follow easily predictable asymmetric cycles akin to *Edgeworth cycles* (Edgeworth 1925): price

increases are fast and large (relenting phase), and are followed by a sequence of small decreases (undercutting phase). The existence of these cycles and the fact that they differ from asymmetric pass-through was first documented by Castanias and Johnson (1993) in US markets, and later by Eckert (2002) and Noel (2007a) in Canada. See also Noel (2009) for a discussion of incomplete pass-through in markets with Edgeworth cycles.

Several authors have documented interesting connections between the structure of local markets and the cycles' attributes. Noel (2007a, b) in particular showed that competitive markets tend to have short-lived cycles (or no cycle at all), while small isolated markets often exhibit long periods of price stability and month-long cycles.

The causes and welfare consequences of these cycles are not well understood. Most authors have rationalised their presence using the dynamic pricing game of Maskin and Tirole (1988), which shows that an *Edgeworth cycle* can emerge as a non-cooperative Markov-perfect equilibrium.

The model matches well many features of gasoline price cycles (see Eckert (2003) and Noel (2007b)). On the other hand, two key assumptions – (i) consumers react instantaneously to price differences and (ii) firms are unable to simultaneously adjust their prices – are at odds with several features of gasoline markets. In particular, it is widely acknowledged that retailers are able to adjust their prices frequently, while consumers are thought to be less informed than firms. Hosken et al. (2008) and Lewis (2008) provide comprehensive analysis of gasoline price dispersions and dynamic price adjustments in US markets.

Moreover, two recent price-fixing cases discovered in Australia and Canada revealed that some features of Edgeworth cycles can be explicitly coordinated by retailers (see Wang (2008), and Clark and Houde (2010)). Related to this, Wang (2009) provides an interesting analysis of an Australian policy that forces retailers to simultaneously change their prices only once a day. Consistent with explicit collusion, the results show that price increases are heavily coordinated when firms can freely adjust their prices. After the policy change, however, although prices still follow Edgeworth-like patterns, firms behave according to a mixed strategy consistent with a war of attrition game.

Evaluation of Market Power

Despite the fact that gasoline is a homogenous product for which prices are easily observed, many authors have argued that firms are able to exert a certain degree of market power.

A first group of papers have looked into the market power assumption by testing for price discrimination. Two examples are particularly interesting. Borenstein (1991) uses the slow decrease in the supply of leaded gasoline during the 1980s to show that retailers selling leaded gasoline were increasingly able to extract rents from consumers as the number of available options shrunk.

Shepard (1991) tests for price discrimination by comparing prices at self- and full-service stations. Her sample includes of a significant fraction of stores offering both types of product, which helps to isolate price discrimination motives from unobserved cost differences. By comparing prices at mixed and traditional stations, she tests the hypothesis that a multi-product firm with market power is better able to price-discriminate than a single-product firm. Her results are conclusive: retail prices for full-service gasoline are significantly higher at multi-product stations. Interestingly, multi-product stations also tend to distort full-service margins more than self-service margins, consistent with a model of second-degree price discrimination.

A second group of authors have studied retail margins from a tacit collusion perspective. Slade (1987, 1992) studied the behaviour of retailers in Vancouver during price wars and developed a test discriminating between a static pricing model and *supergame* strategies. Her results easily reject the static pricing model and uncover interesting asymmetries between major and independent retailers. Majors are acting as price leaders responsible for coordinating price increases, while independents are more likely to initiate price wars.

Borenstein and Shepard (1996) provide a more indirect test, looking at the dynamics of margins and demand. They show that gasoline margins respond positively to expected future demand, consistently with the "price wars during booms" prediction of Rotemberg and Saloner (1986).

A third line of research uses observed mergers to quantify the market power of vertically integrated chains. Hastings (2004) is a leading example of this approach (see also Hastings and Gilbert (2005), Simpson and Taylor (2008) and Taylor et al. (2007)). Her identification argument relies on the idea that retail mergers that are negotiated nationally create sharp changes in market structure that are exogenous to local market conditions. Her results suggest that stores competing directly with independent retailers post significantly lower prices, consistent with vertical restraints and brand loyalty interpretations.

More recently, a few authors have used a structural approach to quantify market power in gasoline markets. Houde (2009) estimates an empirical model of spatial differentiation that incorporates the fact that consumers are mobile within the market, unlike the classic address model proposed by Hotelling (1929). The distance between consumers and firms is defined as the time deviation from home-to-work commuting paths, and elasticities of substitution are directly related to the road network structure and traffic flows (rather than physical distance alone). The results indicate that firms enjoy relatively little market power, especially compared to a model in which consumers are located at a single point. Moreover, retail margins would decrease by about 7% if stores were setting their prices independently.

Hastings (2008) uses a similar approach, but focuses on the availability of rich store-level data on wholesale prices. She first documents that stores are paying different wholesale prices, despite being part of the same brand network. She then simulates a uniform wholesale price regulation. The results suggest that wholesale price discrimination has a pro-competitive effect in this market.

Regulation

Gasoline markets often operate under constraining regulations. Two classes of policies have frequently been discussed: below-cost sales and divorcement regulations.

Below-cost sales regulations are currently in place in nine US states and three Canadian province and the debate is ongoing in many jurisdictions. In some cases below-cost regulations date as far back as the great depression, where many state governments institute "fair-business practice" laws applied to all retail markets. The advocates of these policies typically associate aggressive pricing with predatory and loss-leader strategies. Antitrust authorities typically view such legislations as unnecessary, and they point out that state governments may be too easily convinced by accusations of predation made by lobbying groups representing non-integrated chains of gasoline stations.

Several researchers have evaluated the impact of those policies on prices using cross-sectional data and found significant price increases (see in particular Fenili and Lane (1985), Anderson and Johnson (1999), and Johnson and Romeo (2000)). Recently, Carranza et al. (2009) re-examined this question using a store-level panel dataset in Canada. They found that a policy change in the province of Québec led to a long-run decrease in margins and station productivity that can be largely explained by endogenous changes in the composition of markets. Similar results are also found by Skidmore et al. (2005).

Divorcement acts have been implemented in six states to prevent the vertical integration of major oil refiners in the retail sector. They have typically been justified by theories pointing to anticompetitive motives for vertical integration. However, most empirical papers evaluating those policies suggest that banning vertical integration in general leads to higher prices. See in particular the papers by Barron and Umbeck (1984) and Blass and Carlton (2001). Shepard (1993) and Slade (1996) also provide interesting analysis of contractual arrangements in gasoline markets, which suggest that vertical integration might be more efficient than separation in some context.

Finally, a small number of researchers have recently studied the impact of environmental regulations on the organisation and performance of gasoline markets. Brown et al. (2008) and Muehlegger (2006), for instance, look at the impact of gasoline content regulations on wholesale prices, using reduced-form and structural methodologies respectively. Both found that the imposition of heterogenous content regulations across different geographic markets led to greater market segmentation and higher prices. Ying et al. (2010) compare the public and private provision of insurance for hazard risks of gasoline underground storage tanks. Insurance premiums under public systems are typically uniform and provide little incentive for station owners to upgrade their technology, potentially creating significant moral hazard risks. Yin et al. indeed found that Michigan's transition to private market liability insurance led to a 20% decline in accidents relative to adjacent states providing public insurance coverage (i.e. Illinois and Indiana).

Conclusion

As the previous discussion has demonstrated, gasoline markets have generated an impressive amount of academic research. Many topics are still open and richer datasets are increasingly becoming available, which should lead to even more analysis.

It is fair to say that the literature on market power and dynamic pricing should be better integrated. In particular, we learned from the former that consumers are heterogeneous along several dimensions (e.g. income, commuting, information), which leads to price differences across markets and stores. This heterogeneity could play a role in explaining the causes and consequences of asymmetric pass-through and price cycles, but the dynamic pricing literature has largely ignored the consumer side of the market from their analysis.

Similarly, structural and reduced-form methods have been used to document and measure the importance of differentiation and vertical contracts in explaining profit margins. However, little research has been conducted to study firms' decisions to differentiate themselves (e.g. retail network configuration, choice of amenities, etc.), and design vertical contracts. Large and precise datasets are now available concerning the location of stores and road networks, allowing researchers to study models of product differentiation empirically. Similar information on the terms of contracts between franchisees and upstream suppliers is much harder to obtain, but would be crucial for better understand the role of vertical restraints in this market.

Finally, although this article has focused mainly on the downstream segment of the petroleum industry, it should be noted that much less research has been devoted to studying market power in the upstream segments. In many respects the refinery sector is analogous to the retail sector but is significantly more concentrated. Even so, we know relatively little about the impact of upstream market structure (e.g. localisation and ownership of refineries) on market outcomes such as wholesale prices and capacity utilisation.

See Also

- ▶ Market Power and Collusion in Laboratory Markets
- ▶ Vertical Integration

Bibliography

Anderson, R., and R. Johnson. 1999. Antitrust and sales-below-cost laws: The case of retail gasoline. *Review of Industrial Organization* 14(3): 189–204.

Bacon, R. 1991. Rockets and feathers: The asymmetric speed of adjustment of U.K. retail gasoline prices to cost changes. *Energy Economics* XIII: 211–218.

Barron, J., and J. Umbeck. 1984. The effects of different contractual arrangements: The case of the retail of gasoline market. *Journal of Law & Economics* 27: 313–328.

Blass, A., and D.W. Carlton. 2001. The choice of organizational form in gasoline the choice of organizational form in gasoline retailing and the cost of laws that limit that choice. *Journal of Law and Economics* 44.

Borenstein, S. 1991. Selling costs and switching costs: Explaining retail gasoline margins. *The Rand Journal of Economics* 22(3): 354–369.

Borenstein, S., and A. Shepard. 1996. Dynamic pricing in retail gasoline markets. *The Rand Journal of Economics* 27(3): 429–451.

Borenstein, S., and A. Shepard. 2002. Sticky prices, inventories, and market power in wholesale gasoline markets. *RAND Journal of Economics* 33(1): 116–139.

Borenstein, S., A.C. Cameron, and R. Gilbert. 1997. Do gasoline prices respond asymmetrically to crude oil price changes? *The Quarterly Journal of Economics* 112(1): 305–339.

Brown, J., J. Hastings, E.T. Mansur, and S.B. Villas-Boas. 2008. Reformulating competition? Gasoline content regulation and wholesale gasoline prices. *Journal of Environmental economics and management*.

Carranza, J.E., R. Clark, and J.-F. Houde. 2009, April. Price controls and competition in gasoline retail markets. Working paper, University of Wisconsin-Madison.

Castanias, R., and H. Johnson. 1993. Gas wars: Retail gasoline price fluctuations. *Review of Economics and Statistics* 75: 171–174.

Clark, R., and J.-F. Houde. 2010, June. *Collusion between asymmetric retailers: Evidence from a gasoline price-fixing case*. Mimeo, University of Wisconsin-Madison.

Deltas, G. 2008. Retail gasoline price dynamics and local market power. *The Journal of Industrial Economics* 61: 613–628.

Eckert, A. 2002. Retail price cycles and response asymmetry. *Canadian Journal of Economics* 35: 52–77.

Eckert, A. 2003. Retail price cycles and presence of small firms. *International Journal of Industrial Organization* 21: 151–170.

Edgeworth, F. 1925. The pure theory of monopoly. *Papers Relating to Political Economy* 1: 111–142.

Fenili, R., and W. Lane. 1985. Thou shalt not cut prices. *Regulation* 9(5): 31–35.

Gicheva, D., J. Hastings, and S.B. Villas-Boas. 2007. Revisiting the income effect: Gasoline prices and grocery purchases. Working paper 13614, NBER.

Hastings, J. 2004. Vertical relationships and competition in retail gasoline markets: Empirical evidence from contract changes in southern California. *American Economic Review.*

Hastings, J. 2008, September. Wholesale price discrimination and regulation: Implications for retail gasoline prices. Working paper, Yale University.

Hastings, J., and R. Gilbert. 2005, December. Market power, vertical integration, and the wholesale price of gasoline. *Journal of Industrial Economics*.

Hosken, D., R.S. McMillan, and C. Taylor. 2008. Retail gasoline pricing: What do we know? *International Journal of Industrial Organization* 26(6): 1425–1436.

Hotelling, H. 1929. Stability in competition. *Economic Journal* 39(153): 41–57.

Houde, J.-F. 2009, June. Spatial differentiation and vertical contracts in retail markets for gasoline. Working paper, University of Wisconsin-Madison.

Johnson, R.N., and C.J. Romeo. 2000. The impact of self-service bans in the retail gasoline market. *The Review of Economic and Statistics* 82(4): 625–633.

Levin, C. 2002, May. Gas prices: How are they really set? Hearings before the subcommitte of investigations. US Senate.

Lewis, M. 2005, July. Asymmetric price adjustment and consumer search: An examination of the retail gasoline market. Working paper, Ohio State University. *Journal of Economic Management and Strategy* (Forthcoming).

Lewis, M. 2008. Price dispersion and competition with differentiated sellers. *Journal of Industrial Economics* 56(3): 654–678.

Maskin, E., and J. Tirole. 1988. A theory of dynamic oligopoly, ii: Price competition, kinked demand curves, and edgeworth cycles. *Econometrica: Journal of the Econometric Society* 56(3): 571–599.

Muehlegger, E.J. 2006. *Gasoline price spikes and regional gasoline content regulations: A structural approach*.

Noel, M. 2007a. Edgeworth price cycles, cost-based pricing and sticky pricing in retail gasoline retail markets. *Review of Economics and Statistics* 89(2): 324–334.

Noel, M. 2007b. Edgeworth price cycles: Evidence from the toronto retail gasoline market. *Journal of Industrial Economics* LV(1): 69–92.

Noel, M. 2009. Do gasoline prices respond asymmetrically to cost shocks? The effect of edge-worth cycles. *RAND Journal of Economics* 40(3): 582–595.

Peltzman, S. 2000. Prices rise faster than they fall. *Journal of Political Economy* 108(3): 466–502.

Rotemberg, J.J., and G. Saloner. 1986. A supergame-theoretic model of price wars during booms. *The American Economic Review* 76(3): 390–407.

Shepard, A. 1991. Price discrimination and retail configuration. *Journal of Political Economy* 99(1): 30–53.

Shepard, A. 1993. Contractual form, retail price, and asset characteristics in gasoline retailing. *The Rand Journal of Economics* 24(1): 58–77.

Simpson, J., and C.T. Taylor. 2008. Do gasoline mergers affect consumer prices? The Marathon Ashland Petroleum and Ultramar Diamond Shamrock transaction. *The Journal of Law and Economics* 51(1): 135–152.

Skidmore, M., J. Peltier, and J. Alm. 2005. Do state motor fuel sales-below-cost laws lower prices? *Journal of Urban Economics* 57(1): 189–211.

Slade, M.E. 1987. Interfirm rivalry in a repeated game: An empirical test of tacit collusion. *The Journal of Industrial Economics* 35(4, The Empirical Renaissance in Industrial Economics): 499–516.

Slade, M.E. 1992. Vancouver's gasoline-price wars: An empirical exercise in uncovering supergame strategies. *The Review of Economic Studies* 59(2): 257–276.

Slade, M.E. 1996. Multitask agency and contract choice: An empirical exploration. *International Economic Review* 37(2): 465–486.

Tappata, M. 2009. Rockets and feathers. understanding asymmetric pricing. *RAND Journal of Economics* 40(4).

Taylor, C.T., N. Kreisle, and P.R. Zimmerman. 2007, September, Forthcoming. Vertical relationships and competition in retail gasoline markets: Comment. AER.

Wang, Z. 2008. Collusive communication and pricing coordination in a retail gasoline market. *Review of Industrial Organization* 32(1): 35–52.

Wang, Z. 2009. (mixed) strategy in oligopoly pricing: Evidence from gasoline price cycles before and under a timing regulation. *Journal of Political Economy* 117(6).
Yang, H., and L. Ye. 2008. Search with learning: Understanding asymmetric price adjustments. *RAND Journal of Economics* 39: 547–564.
Ying, H., H. Kunreuther, and M. White. 2010. Risk-based pricing and risk-reducing effort: Does the private insurance market reduce environmental accidents? *Journal of Law & Economics* (Forthcoming).

Gauge Functions

Peter Newman

Consider the standard two-product diagram which depicts an opportunity set P with production frontier fr(P). For any point x^1 inside P it would be useful to have a measure of just how inefficient it is, i.e. to gauge how far it is from the frontier. A simple way of doing this is, first to find that point $\bar{x} \in \text{fr}(P)$ which is just a scale change of x^1, so that $x^1 = \lambda_1 \bar{x}$ for some $\lambda_1 \in [0, 1)$. Then a function $J(. \mid P)$ that calibrates any such point with respect to P is defined by putting $J(x^1 \mid P) = \lambda_1$. For this to be a sensible measure of efficiency, it should obviously have the property that $J(x \mid P) = 1$ if and only if (iff) $x \in \text{fr}(P)$.

Similarly, for any point x^2 outside P it can be asked: How much would productive capacity (i.e. P) have to grow in order that x^2 be producible? Again, a simple measure of this would be to find $\ddot{x} \in \text{fr}(P)$ such that $x^2 = \lambda_2 \ddot{x}$ for some $\lambda_2 > 1$, and put $J(x^2 \mid P) = \lambda_2$. Thus $J(x \mid P)$ becomes a general measure of the producibility of x with respect to P.

In the same way, in the theory of consumer's preferences consider a better set $B^t = \{x \in R^n: x \gtrsim x^t\}$ for some 'target' bundle x^t, and let x^3 be any bundle that lies above the indifference surface I^t which bounds B^t from below. An obvious measure of how much x^3 could be reduced and the resulting bundle still remain in B^t is given by finding $\bar{x} \in I^t$ such that $x^3 = \mu_3 \bar{x}$ for some $\mu_3 > 1$. Provided $j(x \mid B^t) = 1$ iff $x \in I^t$, μ_3 is then a measure of the redundancy of x^3 in achieving the target level of satisfaction represented by x^t. Again, if x^4 lies below I^t and $x^4 = \mu_4 \ddot{x}$ for some $\ddot{x} \in I^t$ and $\mu_4 \in [0, 1)$, then μ_4 is a measure of the shortfall of x^4 in achieving x^t. In each case, putting $j(x \mid B^t) = \mu$ defines a function that gauges the performance of the actual bundle x with respect to the set B^t, hence to the target bundle x^t.

These two functions $J(. \mid P)$ and $j(. \mid B^t)$ are examples of what in this essay will be called *gauge* and *s-gauge* functions, respectively. Such functions form the basis of one of the two main duality schemes in economics, the other scheme being that of Fenchel transforms (for which see "▶ Duality").

History

As already indicated, gauge functions are of two types. The first, simply called *gauges*, are a direct generalization of Minkowski's Distanzfunktion (e.g. (1911)) and so are sometimes referred to as *Minkowski functionals*. Functions of this type are often used in mathematics but are as yet rarely employed (at least explicitly) in economics. They are best suited for bounded sets lying near the origin, such as P above and the trading sets X_i of McKenzie (1981, p. 820).

The second type of gauge function is almost unknown in mathematics but often used in economics, chiefly for unbounded sets that do not contain the origin, such as B^t above and analogous sets in the theory of production. Economists have given these functions many names, among them 'distance', 'transformation' and 'deflation' function. The name *s-gauges* used here both pays homage to Shephard (1953) and emphasizes their affinity with gauges.

S-gauges were introduced more or less simultaneously by Debreu (1951), Malmquist (1953) and Shephard (1953), each in his own way. Debreu, concerned with general equilibrium, defined his function 'the coefficient of resource utilization' for the better set in commodity space whose lower boundary is a Scitovsky community indifference surface; it was obtained as the solution to an optimization problem and is actually the

inverse of an s-gauge. Malmquist, concerned with index numbers and hampered by misprints, defined a 'quantity index' (pp. 230–32) as s-gauges of better sets B^t, and a 'price index' (pp. 213–15) as s-gauges of the sets B_0^t (see Section "S-Polar Sets") that are dual to B^t.

Shephard, concerned with production functions, defined s-gauges for surfaces in the strictly positive orthant and showed that cost functions could be regarded as s-gauges of dual surfaces in the price space. Oddly, he appeared to regard s-gauges as an example of Minkowski's Distanzfunktion (1953, pp. 6), even though the latter is a convex function defined only for convex compact sets with the origin as interior point, while Shephard's function is concave and defined only for convex unbounded sets that clearly do not contain the origin. Later he implicitly recognized this anomaly by referring to s-gauges as 'an adaptation ... of the Minkowski distance function' (1970, p. 66).

Notwithstanding, his discussion was by far the most comprehensive of the three and fully warrants Diewert's judgement (1982, p. 551): 'the first modern, rigorous treatment of duality theory' – at least in economics.

After this pioneering work there was a long gap, until s-gauges reappeared in the 1970s with the work (in alphabetical order) of Blackorby, Primont and Russell (e.g. 1978), Deaton (1979), Deaton and Muellbauer (1980), Diewert (e.g. 1982, which contains a full bibliography), Gorman (1970, 1976), Hanoch (1978), Jacobsen (1972), McFadden (1978), Ruys (1972), Ruys and Weddepohl (1979), Shephard (1970) and Weddepohl (1972). It is sometimes argued that Wold (1943) and Uzawa (1964) were early contributors to this literature, but Wold used only a Euclidean norm and explicitly rejected a Minkowskian approach (1943, pp. 231–2), while Uzawa's paper was actually a nice formulation of the 1–1 correspondence between closed convex sets and their support functions.

Although these later investigations showed a wide range of application for s-gauges, apart from Weddepohl's contribution their formal theory remains more or less as Shephard left it in 1953, still seriously incomplete. For example, economists usually define s-gauges with respect to utility and production *functions* rather than sets (which was the original Minkowskian tradition) and partly for this reason place severe and unexplained restrictions on the relevant domains. Moreover, their discussions are typically confined to convex preferences and technologies and to finite-dimensional spaces.

This essay sets out a formal and coherent account of gauges and s-gauges taken together, although no proofs are given. The primal (e.g. quantity) and dual (e.g. price) spaces will be denoted X and Y, respectively. Though the non-specialist reader need only consider the familiar space $X = R^n = Y$, the natural frame of reference of this theory is where $\langle X, Y\rangle$ are a pair of (in general) infinite-dimensional Hausdorff topological vector spaces 'in duality' (see e.g. Robertson and Robertson 1964), a class which for example includes Banach spaces and their duals.

Bonnesen and Fenchel (1934) give a detailed discussion of and references to Minkowski's original Distanzfunktion on R^n; Cassels (1959) is also useful. Discussions of gauges (Minkowski functionals) in more general spaces are to be found in works on functional analysis, e.g. Bourbaki (1953), Kothe (1969) and Moreau (1967). For mathematical concepts not explained here, see the entries on "▶ Convex Programming" and "▶ Duality".

Definitions and Simple Properties

Gauges

A set $C \subset X$ is a *cone with vertex 0* if $x \in C \Rightarrow \lambda x \in C$ for all $\lambda > 0$. Deconstructing this idea into two others, *starred* and *haloed* sets, the *starred hull* $e(a)$ of any set $A \subset X$ is given by $e(A) = \{v \text{ e } X{:}v = \lambda x,\ x \in A\ 0 \leq \lambda \leq 1\}$ and the *haloed hull* $h(A)$ by $h(A) = \{w \in X{:}w = \lambda x,\ x \in A, \lambda \leq 1\}$. Then A is *starred* iff $A = e(A)$ and *haloed* iff $A = h(A)$. The *conical hull* $C(A)$ is $e(A) \cup h(A)$, and A is a cone with vertex *0* iff $A = C(A)$.

For any $A \subset X$ its *gauge* is the numerical function $J(.|\ A)$: $X \to [0, \infty]$ defined by

$$\begin{aligned} \boldsymbol{J}(x|A) &= \inf\{\lambda > 0 : x \in A\} \quad \text{if } x \neq 0 \\ &= 0 \quad \text{if } x = 0 \end{aligned} \tag{1}$$

In 'seeing' this definition, think A as rather like the unit sphere, or at least as being near the origin and possibly bounded. If $x \notin A$ the set is to be enlarged by the magnification factor $\lambda > 1$ until it just engulfs x, the corresponding value of λ then being $J(x \mid A)$, while if $x \in A$ then A is to be uniformly shrunk by the contraction factor $\lambda \leq 1$ until it is on the verge of parting company with x.

It follows from (1) that $J(.\,|A)$ is always proper, and positively homogeneous (ph), (i.e. $J(\lambda x \mid A)) = \lambda J(x \mid A)$ for all $\lambda \geq 0$. More surprisingly, $J(.\,|e(A)) \equiv J(.\,|A)$, so that it is natural for the theory of gauges to deal only with starred sets. Notice that if $A = C(A)$ then $J(.\,|A) \equiv \delta(.\,|A)$, where the latter function is the *indicator* of A, i.e. $\delta(x \mid A) = 0$ if $x \in A$, and $= \infty$ otherwise.

If A is starred and (topologically) closed, then $J(.\mid A)$ is lower semicontinuous (lsc) on the whole of X, and decreasing in A (i.e. $A^1 \subset A^2$ iff $J(.\,|A^1) \geq J(.\,|A^2)$), while $A = \{x \in X : J(x \mid A) \leq 1\}$. Convex sets that contain the origin (and so are starred) have gauges that are not only ph but also *subadditive* ($J(v + w \mid A) \leq J(v \mid A) + J(w \mid A)$ for all $v, w \in X$), and so they are convex functions. Finally, if A is convex, closed and such that 0 is in the topological interior (int) of A, then the (topological) boundary or *frontier* of A is $\{x \in X : J(x \mid A) = 1\}$.

As already noted, Minkowski's Distanzfunktion was defined only for convex compact $A \subset R^n$ such that $0 \in \text{int}\, A$, whereas its generalization $J(.\mid A)$ can be used for much wider classes of both sets and spaces. However, in mathematics gauges are in fact typically limited to sets that are convex, closed, *balanced* (if $|\lambda| \leq 1$ and $x \in A$ then $\lambda x \in A$) and such that $0 \in \text{int}\, A$, i.e. sets that are very much like the unit sphere.

S-Gauges

The formal treatment of s-gauges in mathematics seems confined to Phelps (1963), although economists have made some valuable contributions, in particular Shephard (1953) and Weddepohl (1972). The next definition is best 'seen' if B is taken to be haloed (hence unbounded) and *not* containing the origin.

For any $B \subset X$ its *s-gauge* is the numerical function $j(.\mid B): X \to \{-\infty\} \cup [0, \infty]$ defined by

$$\begin{aligned} j(x|B) &= \sup\{\mu > 0 : x \in \lambda B\} \quad \text{if } x \neq 0 \\ &= 0 \quad \text{if } x = 0 \end{aligned} \tag{2}$$

If $x \notin B$, $j(x \mid B)$ is found by pulling B uniformly towards the origin via the contraction factor $\mu \in (0, 1]$ until μB just reaches x. If $x \in B$, then $j(x \mid B)$ is found by making B recede radially away from the origin through multiplication by the expansion factor $\mu \geq 1$ until μB is on the verge of leaving x behind. Notice that $j(x \mid B) = 0$ iff $x = 0$, unlike the situation with gauges.

A point $v \in B$ is *internal to B* if for any other $x \in X$ there exists $\varepsilon > 0$ such that $(v + \lambda x) \in B$ if for all λ with $|\lambda| < \varepsilon$; this is an algebraic rather than topological idea of what it means to be inside a set. If the origin 0 is one of B's internal points, so that every point x in X can as it were be drawn into B by scaling that point down by a suitable contraction factor λ, then B is *absorbent*. Denote by $X \setminus B$ the set-theoretic difference X *less* B, and put $c(B) = C(B) \setminus \{0\}$.

It follows from (2) that $j(.\,|B)$ is ph and that it is n-proper if $X \setminus B$ is absorbent. Phelps (1963, Prop, 2iii) proved that $j(.\,|h(B)) \equiv j(.\,|B)$, so it is natural to assume that B is always haloed. If B is haloed and closed, then $j(.|B)$ is increasing in B, i.e. $B^1 \subset B^2$ iff $j(.\,|B^1) \leq j(.\,|B^2)$, and $B = \{x \in X : j(x \mid B) \geq 1\}$. If B is closed and $0 \notin B$, then $j(.\mid B)$ is upper semicontinuous (usc) on $C(B)$. If B is convex, haloed and such that $X \setminus B$ is absorbent, then $j(.\,|B)$ is superadditive ($j(v + w \mid B) \geq j(v \mid B) + j(w \mid B)$ for all $v, w \in X$), and so concave. Finally, if B is haloed, convex, closed and such that $\text{int}\, B \neq \emptyset$, then the frontier of B is $\{x \in X : j(x \mid B) = 1\}$.

These properties of s-gauges should be compared with the corresponding properties of gauges.

Dual Sets

Polar Sets

For any $A \subset X$ its *polar set* $A^0 \subset Y$ is defined by

$$A^0 = \{y \in Y : \langle x, y\rangle \leqslant 1 \quad \text{for all } x \in A\} \tag{3}$$

where $\langle x, y\rangle$ denotes the value of the linear functional y at x (if $X = R^n = Y$ then it is the inner product of x and y). If $A = C(A)$ then it is easy to check that A^0 coincides with the *polar cone* of A ($= \{y \in Y: \langle x, y\rangle \leq 0$ for all $x \in A\}$), whence the name of this generalization.

Let $\Gamma^0(Y)$ denote the set of all those subsets of Y that are convex, closed and contain the origin (hence are starred); similarly for $\Gamma^0(X)$. From (3) each y in A^0 satisfies a family of weak linear inequalities (one for each x in A), so that A^0 is convex and closed; moreover, obviously $0 \in A^0$ for any A. Hence $A^0 \in \Gamma^0(Y)$.

The *bipolar set* $A^{00} \subset X$ of any $A \subset X$ is given by

$$A^{00} = \{x \in X : \langle x, y\rangle \leqslant 1 \quad \text{for all} \quad y \in A^0\}. \quad (4)$$

One can define $A^{000}, A^{0000}, \ldots$ in ways analogous to (3) and (4) respectively, but it is easily shown that $A^{000} = A^0, A^{0000} = A^{00} \ldots$ etc. Other simple properties are that $A \subset A^{00}$, and that $A_1 \subset A_2$ implies $A_2^0 \subset A_1^0$ and $A_1^{00} \subset A_2^{00}$.

Since $A^{00} \in \Gamma^0(X)$ it is clearly necessary for $A = A^{00}$ that A itself be in $\Gamma^0(X)$. The next result, saying that this condition is also sufficient, is the fundamental theorem of the theory of gauges. Discovered by Dieudonné and Schwartz (1950), it is equivalent to the Hahn-Banach Theorem, to 'the' Theorem of the Separating Hyperplane, and to the Fenchel-Moreau theorem (for which see "▶ Duality").

Theorem 1 (*Bipolar Theorem*). For any $A \subset X$, $A^{00} = K(e(A))$.

Here, for any set $M \subset X$, $K(M)$ denotes its *closed convex hull*, i.e. the intersection of all the convex closed sets that contain M. An equivalent version of Theorem 1 is that $A^{00} = K(A \cup \{0\})$, so that for example McKenzie's $\overline{X}_i$ (1981, p. 825) is simply $X^{00}{}_i$.

S-Polar Sets

For any $B \subset X$ its *s-polar* set $B_0 \subset Y$ is given by

$$B_0 = \{y \in Y : \langle x, y\rangle \geqslant 1 \quad \text{for all} \quad x \in B\} \quad (5)$$

and its *s-bipolar set* $B_{00} \subset X$ by

$$B_{00} = \{x \in X : \langle x, y\rangle \geqslant 1 \quad \text{for all} \quad y \in B_0\} \quad (6)$$

As before, it can be shown that $B \subset B_{00}$, that $B_{000} = B_0$ and $B_{0000} = B_{00}$ etc., and that $B^1 \subset B^2$ implies $B^2{}_0 \subset B^1{}_0$ and $B^1{}_{00} \subset B^2{}_{00}$. However, it is obvious from (5) that if $0 \in B$ then $B_0 = \emptyset$, which indicates a major asymmetry between polar and s-polar sets.

Denote the class of all non-empty convex closed haloed subsets of Y that do *not* contain the origin by $\Gamma_0(Y)$, and similarly for $\Gamma_0(X)$; since these sets are haloed, they are all unbounded. Although $A^0 \in \Gamma^0(Y)$ and $A^{00} \in \Gamma^0(X)$ for each $A \subset X$, it is *not* true that $B_0 \in \Gamma_0(Y)$ and $B_{00} \in \Gamma_0(X)$ for each $B \in X$. For example, suppose $X = R$ and that $B = \{-b, b\}$ for some number $b \neq 0$. Then $B_0 = \emptyset$ and $B_{00} = R$, though neither $\emptyset$ nor R is in $\Gamma_0(R)$.

Fortunately, by using a separating hyperplane theorem one can obtain precise information about when B_0 exists.

Theorem 2 For any $B \subset X, B_0 = \emptyset$ iff $0 \in K(B)$.

Corollary 1.0 $\notin K(B)$ iff [$B_0 \in \Gamma_0(Y)$ and $B_{00} \in \Gamma_0(X)$].

A basic theorem corresponding to Theorem 1 is

Theorem 3 For any $B \subset X$ such that $0 \notin K(B)$, $B_{00} = K(h(B))$.

A partial converse to this is

Proposition 1 If $K(h(B)) \neq X$, then $B_{00} = K(h(B))$ implies $0 \notin K(B))$.

That the conditional in this result is essential follows from the previous example of $B = \{-b, b\}$, since there $R = B_{00} = K(h(B))$ yet $0 \in K(B))$.

Transforms

Polar Transforms

For any $A \subset X$, the *polar transform* of its gauge J $(\cdot \mid A)$ is the gauge $J(\cdot \mid A^0)$ of its polar set A^0, and the

bipolar transform of $J(\cdot \mid A)$ is the gauge $J(\cdot \mid A^{00})$ of its bipolar set A^{00} (the term is due to Young (1969, p. 108)). Since each of these sets is convex, closed and contains the origin, it follows from earlier results that each of the transforms is convex, ph and lsc on X.

Define the *support function* $S(\cdot \mid A)$: $Y \to [-\infty, \infty]$ of A by

$$S(y|A) = \sup\{\langle x, y\rangle : x \in A\} \tag{7}$$

These functions are ph, lsc iff A is closed, and convex iff A is convex.

The next result is simple to prove but very important.

Theorem 4 For any $A \subset X$,

$$J(\cdot|A^0) \equiv S(\cdot|A^{00}) \text{ and } J(\cdot|A^{00}) \equiv S(\cdot|A^0).$$

While each transform maps to the interval $[0, \infty]$, for several reasons (such as the problem of invertibility) it is important to know when it is actually positive and finite, i.e. maps to $(0, \infty)$.

Proposition 2 For any $A \subset X$,

(a) $[\{y \in Y{:}0 < J(y \mid A^0) < \infty\} = c(A^0)]$ iff $[y \neq 0 \to S(y \mid A^{00}) \neq 0]$
(b) $[\{x \in X{:}0 < J(x \mid A^{00}) < \infty\} = c(A^{00})]$ iff $[x \neq 0 \to S(x \mid A^0) \neq 0]$.

These conditions are precise but restrictive. Moreover, it is not easy to 'see' what A should look like in order that they be satisfied. So it is useful to have

Proposition 3 For any $A \subset X$,

(a) (i) $[\{y \in Y{:}0 < J(y \mid A^0) < \infty\} = c(A^0)] \to [A^{00}$ absorbent]
(ii) $[0 \in \text{int } A^{00}] \to [\{y \in Y{:}0 < J(y \mid A^0) < \infty\} = c(A^{00})]$
(b) $[\{x \in X{:}0 < J(x \mid A^{00}) < \infty\} = c(A^{00})]$ iff [A bounded]
(c) If $X = R^n = Y$,

$$[\{y \in Y : 0 < J(y|A^0) < \infty\} = c(A^0)] \text{ iff } 0 \in \text{int} A^{00}.$$

The last of these results can be deduced from Rådström (1949–50, para. 4, p. 28) or Rockafellar (1970, Cor. 14.5.1, p. 125).

S-Polar Transforms

For any $B \subset X$, the *s-polar transform* of $j(\cdot \mid B)$ is the s-gauge $j(\cdot \mid B_0)$ of B_0, and similarly the *s-bipolar transform* of $j(\cdot \mid B)$ is the s-gauge $j(\cdot \mid B_{00})$ of B_{00}. Provided B_0 exists, then from earlier results each s-polar transform is concave, ph and use on $C(B_0)$ (or $C(B_{00})$, as the case may be). Unlike the case of polar transforms, the positivity and finiteness of s-polar transforms offers no difficulty, as shown by

Proposition 4 For any $B \subset X$,

(a) $\{y \in Y{:}0 < j(y \mid B_0) < \infty\} = c(B_0)$
(b) $[\{x \in X{:}0 < j(x \mid B_{00}) < \infty\} = c(B_{00})]$ iff $B_0 \neq \emptyset$

As an application of (b), suppose that B is a better set for some target x^t and that $B \in \Gamma_0(X)$, so that $B_0 \neq \emptyset$ and $B_{00} = B$. If the indifference surface I^t is asymptotic to each axis then $j(x \mid B)$ is positive and finite for each strictly positive bundle x. This is a rationale for the common requirement that 'distance functions' be defined only on the strictly positive orthant.

For s-polar transforms it is not support functions that are relevant but *concave support functions* $s(\cdot \mid B)$: $Y \to [-\infty, \infty]$ of B, defined by

$$S(y|B) = \inf\{\langle x, y\rangle : x \in B\} \tag{8}$$

Such functions are ph, use iff B is closed and concave iff B is convex. Unfortunately, the simple universality of Theorem 4 is not available for s-polar transforms. Without further assumption, the best that can be done is

Proposition 5 For any $B \subset X$,

(a) $y \in c(B_0) \to j(y \mid B_0) = s(y \mid B_{00})$
(b) $[x \in c(B_{00}) \to j(x \mid B_{00}) = s(x \mid B_0)]$ iff $B_0 \neq \emptyset$.

In looking for conditions to strengthen these results a hint is provided by the fact, noted earlier, that s-gauges vanish *only* at the origin. It follows that to have identity between s-polar transforms and concave support functions the latter must have that property as well. Now Arrow's famous 'exceptional case' (1951, pp. 527–8) referred precisely to a concave support function (i.e. the cost function for Individual 2's better set) which vanished for some nonzero price vector. This motivates the following

Definition A set $M \subset X$ such that $y \neq 0 \rightarrow s(y \mid M) \neq 0$ is called *Arrovian*, and similarly for any $N \subset Y$ such that $x \neq 0 \rightarrow s(x \mid N) \neq 0$.

Geometrically, a non-empty $M \subset X$ is Arrovian iff none of the affine hyperplanes that support it from below passes through the origin. Economically, the idea is closely related to the existence of the 'locally cheaper points' of McKenzie (1957). That it is just the strengthening condition required is shown by

Theorem 5 For any $B \subset X$:

(a) $j(\cdot \mid B_0) \equiv s(\cdot \mid B_{00})$ iff B_{00} is Arrovian.
(b) $[j(\cdot \mid B_{00}) \equiv s(\cdot \mid B_0)$ iff B_0 is Arrovian] iff $B_0 \neq \emptyset$.

Comparison Between Polar and S-Gauge Transforms

The assumption that a set be Arrovian is clearly an s-gauge version of the corresponding assumptions for $S(\cdot \mid A^{00})$ and $S(\cdot \mid A^{0})$ that were prescribed in Propositions 2(a) and 2(b), respectively. However, Proposition 3 showed the latter to be closely related (if not equivalent) to more intuitively understandable assumptions, namely, absorbency and boundedness. Hence there was no need for a separate labelling of 'gauge-Arrovian' sets. For s-gauges, however, there appear to be no such simple characterizations of the Arrovian property.

It follows from these considerations that, in any given application, properties which at first sight look very different may simply be 'gauge/s-gauge' versions of the same idea. Suppose $A \in \Gamma^0(R^n)$ and $B \in \Gamma_0(R^n)$ so that $A^{00} = A$, $B_0 \neq \emptyset$ and $B_{00} = B$. Then from Propositions 2 and 3 the gauge version of the s-gauge condition that B be Arrovian (e.g. that it always has locally cheaper points) is that $0 \in \operatorname{int} A$, while the gauge version of B_0 being Arrovian is that A be bounded. Thus different appearances may mask similar roles, forming disguises which are not easily penetrated without the help of duality theory.

Mahler's Inequality

Gauges

Polar and bipolar transforms of gauges satisfy a fundamental inequality, which plays the same role in the theory of gauges that (W.H.) Young's Inequality plays in the theory of Fenchel transforms (see "▶ Duality"). L.C. Young (1939) attributed the statement and proof (for R^n) of this inequality of Mahler (1939), though saying also (p. 569) that 'it is implicitly contained in the work of Minkowski'. An inkling of its importance is that the celebrated Cauchy-Schwarz Inequality is a special case (see "▶ Inequalities").

Theorem 6 (Mahler's Inequality). For any $A \subset X$.

$$\forall x \in X, \quad \forall y \in Y \quad J\left(x|A^{00}\right)J\left(y|A^{0}\right) \geqslant \langle x, y\rangle \tag{9}$$

Those pairs $(x\, y)$ such that equality holds in (9) may be called *polar* to each other. Notice that if $A \in \Gamma^0(X)$ then, from Theorems 1 and 4, (9) is equivalent to

$$\forall x \in X, \quad \forall y \in Y \quad J(x|A)J(y|A) \geqslant \langle x, y\rangle \tag{10}$$

In economic applications, if x is a quantity vector and y a price vector then in (10) $J(\cdot \mid A)$ is dimensionless and $S(\cdot \mid A)$ has dimension 'price × quantity'. New characterizations of the polar and bipolar transforms are provided by

Proposition 6 For any $A \subset X$:

(a) If $0 \in \operatorname{int} A^{00}$, $\forall x \in X$

$$J\left(x|A^{00}\right) = \sup\left\{\langle x, y\rangle : y \in Y, J\left(y|A^{0}\right) = 1\right\} \tag{11}$$

(b) If A is bounded, $\forall y \in Y$

$$J(y|A^0) = \sup\{\langle x, y\rangle : x \in X,\ J(x|A^{00}) = 1\} \tag{12}$$

If $A \in \Gamma^0(X)$ then (11) and (12) simplify considerably, just as (10) simplifies (9).

S-Gauges

The s-gauge version of Mahler's Inequality, though still remarkable enough, is not so unrestricted as that for gauges.

Theorem 7 For any $B \subset X$ such that $0 \notin K(B)$:

(a) If B_{00} is Arrovian, $\forall x \in X, \forall y \in C(B_0)$

$$j(x|B_{00})j(y|B_0) \leqslant \langle x, y\rangle \tag{13}$$

(b) If B_0 is Arrovian, $\forall x \in C(B_{00}), \forall y \in Y$

$$j(x|B_{00})j(y|B_0) \leqslant \langle x, y\rangle \tag{14}$$

The characterizations of the s-polar and s-bipolar transforms which are analogous to (11) and (12) are similarly restricted.

Proposition 7 For any $B \subset X$ such that $0 \notin K(B)$:

(a) If B_{00} is Arrovian, $\forall x \in c(B_{00})$

$$j(x|B_{00}) = \inf\{\langle x, y\rangle : y \in Y,\ j(y|B_0) = 1\} \tag{15}$$

(b) If B_0 is Arrovian, $\forall y \in c(B_0)$

$$j(y|B_0) = \inf\{\langle x, y\rangle : x \in X,\ j(x|B_{00}) = 1\} \tag{16}$$

A close reading of Gorman (1976) and Deaton (1979) will show that Theorem 7 and Proposition 7 formalize and generalize several of the results for s-gauges ('distance functions') obtained in those papers.

Subdifferentials

Sub- and superdifferentials generalize differentials in a way especially appropriate to convex and concave functions, respectively; for a fairly detailed discussion see "▶ Duality" (Section II). In particular, conditions of optimality (e.g. minimization of cost, maximization of profit) are expressible naturally by the non-emptiness of sub- and super differentials. Denote the subdifferential (resp., superdifferential) of the gauge (resp., s-gauge) of any set M by $\partial J(\cdot \mid M)$ (resp., $\Delta j\ (\cdot \mid M)$). Then a comprehensive result for subdifferentials of polar and bipolar transforms is

Theorem 8 For any $A \subset X$:

(a) $y^1 \in \partial J(x^1 \mid A^{00})$ iff $[y^1 \in A^0$ and y^1 achieves $S(x^1 \mid A^0)]$

(b) $x^1 \in \partial J(y^1 \mid A^0)$ iff $[x^1 \in A^{00}$ and x^1 achieves $S(y^1 \mid A^{00})]$

(c) If $y^1 \in c\,(A^0)$ and $0 \in \operatorname{int} A^{00}$, then

$$\begin{aligned}&\left(y^1|J\left(y^1/A^0\right)\right) \in \partial J\left(x^1|A^{00}\right)\\ &\text{iff}\left[J\left(x^1|A^{00}\right)J\left(y^1|A^0\right) = \left\langle x^1, y^1\right\rangle\right]\end{aligned}$$

(d) If $x^1 \in c(A^{00})$ and A is bounded, then

$$\begin{aligned}&\left(x^1|J\left(x^1/A^{00}\right)\right) \in \partial J\left(y^1|A^0\right)\\ &\text{iff}\left[J\left(x^1|A^{00}\right)J\left(y^1|A^0\right) = \left\langle x^1, y^1\right\rangle\right].\end{aligned}$$

The combination of (c) and (d) obviously gives conditions under which the subdifferentials of the polar and bipolar transforms are *inverse* to each other in the sense of set-valued mappings, i.e. when

$$\begin{aligned}&\left(y^1J\left(y^1|A^0\right)\right) \in \partial J\left(x^1|A^{00}\right)\\ &\text{iff}\left(x^1|J\left(x^1|A^{00}\right)\right) \in \partial J\left(y^1|A^0\right).\end{aligned}$$

As one might expect by now, the corresponding results for superdifferentials of s-gauges are slightly more restrictive.

Theorem 9 For any $B \subset X$ such that $0 \notin K(B)$:

(a) If B_0 is Arrovian,

$$y^1 \in \Delta j(x^1|B_{00})$$
$$\text{iff}[y^1 \in B_0 \text{ and } y^1 \text{ achieves } s(x^1|B_0)]$$

(b) If B_{00} is Arrovian,

$$x^1 \in \Delta j(y^1|B_0)$$
$$\text{iff}[x^1 \in B_{00} \text{ and } x^1 \text{ achieves } s?(y^1|B_{00}]$$

(c) If $y^1 \in c(B_0)$ and B_0 is Arrovian, then

$$\left(y^1/j(y^1|B_0) \in \Delta j(x^1|B_{00})\right.$$
$$\text{iff } \left[j(x^1/B_{00})\left(j(y^1|B_0) = \langle x^1, y^1\rangle\right)\right]$$

(d) If $x^1 \in c(B_{00})$ and B_{00} is Arrovian, then

$$\left(x^1/j(x^1|B_{00}) \in \Delta j(y^1|B_o)\right.$$
$$\text{iff } \left[j(x^1|B_{00})j(y^1|B_0) = \langle x^1, y^1\rangle\right]$$

Notice the subtle changes between this theorem and Theorem 8. As already noted, the condition $o \in \text{int } A^{00}$ corresponds to B_{00} being Arrovian. In Theorem 8(c) the first of these conditions is used to assure the meaningfulness of the ratio $y^1/J(y^l \mid B^o)$, while in Theorem 9(d) the second condition is used, not to validate the ratio $x^1/j(x^1 \mid B_{00})$ (this follows from Proposition 4(b)), but to assure that $j(y \mid B00) = s(y \mid B0)$ for all $y \in Y$ (Theorem 5(a)). A similar analysis holds for the conditions that A be bounded (Theorem 8(d)) and that β_0 be Arrovian (Theorem 9(c)).

Observe that Theorem 9(c) and (d) provide conditions under which the superdifferential mappings associated with s-polar and s-bipolar transforms are inverse to each other in the sense of set-valued mappings. Indeed, since superdifferentiability generalizes the differentiability of concave functions, Theorem 9 formalizes and generalizes most of the optimality conditions in the literature on s-gauges, e.g. Properties 5–8 in Deaton (1979, pp. 394–6).

Conclusion

Gauges and their polar transforms, together with s-gauges and their s-polar transforms, constitute a powerful and general duality scheme for economic theory. This scheme has already had many applications to the theory of production and consumption, to index number theory, to optimal taxation, and to the theory of income distribution, to name just a few. Many more can be expected, including applications to general equilibrium theory. The scheme is especially well suited to situations where the appropriate normalization of prices is by total expenditure or income, just as the other main duality scheme – Fenchel transforms – is suitable when the appropriate normalization is by a single numeraire good.

Finally, the close parallelism between the theory of gauges and that of s-gauges suggests the existence of some transformation of variables under which the two theories would be seen simply as two aspects of one unified theory.

See Also

- Cost Minimization and Utility Maximization
- Duality
- Homogeneous and Homothetic Functions
- Rationing

Bibliography

Arrow, K.J. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, ed. J. Neyman, 507–532. Berkeley: University of California Press.

Blackorby, C., D. Primont, and R.R. Russell. 1978. *Duality, separability and functional structure: Theory and economic applications*. New York: Elsevier/North-Holland.

Bonnesen, T., and W. Fenchel. 1934. *Theorie der Konvexen Körper*. Berlin: Springer. Reprinted, New York: Chelsea, 1948.

Bourbaki, N. 1953. *Espaces vectoriels topologiques, Chapitres I-II*, Actualites Scientifiques et Industrielles No. 1189. Paris: Hermann.

Cassels, J.W.S. 1959. *An introduction to the geometry of numbers*. Berlin: Springer.

Deaton, A. 1979. The distance function in consumer behaviour with applications to index numbers and optimal taxation. *Review of Economic Studies* 46: 391–406.
Deaton, A., and J. Muellbauer. 1980. *Economics and consumer behaviour*. Cambridge: Cambridge University Press.
Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19: 273–292.
Dieudonné, J., and L. Schwartz. 1950. La dualité dans les espaces (^) et (+^). *Annales de L'Institut Fourier*, Université de Grenoble, 1: 61–101.
Diewert, W.E. 1982. Duality approaches to microeconomic theory. In *Handbook of mathematical economics*, vol. II, ed. K.J. Arrow and M.D. Intriligator, 535–599. Amsterdam: North-Holland.
Fuss, M., and D. McFadden (eds.). 1978. *Production economics: A dual approach to theory and applications, vol. 1, the theory of production*. Amsterdam: North-Holland.
Gorman, W.M. 1970. *Quasi-separable preferences, costs and technologies*. Chapel Hill: Mimeo, Department of Economics, University of North Carolina.
Gorman, W.M. 1976. Tricks with utility functions. In *Essays in economic analysis*, ed. M.J. Artis and A.R. Nobay. Cambridge: Cambridge University Press.
Hanoch, G. 1978. Symmetric duality and polar production functions, Chapter 1.2, 111–131, in Fuss and McFadden (1978).
Jacobsen, S.E. 1972. On Shephard's duality theorem. *Journal of Economic Theory* 4: 458–464.
Köthe, G. 1969. *Topological vector spaces I*. New York: Springer.
Mahler, K. 1939. Ein Übertragungsprinzip für konvexe Körper. *Casopis pro Pêstováni Matematiky a Fysiky* 63: 93–102.
Malmquist, S. 1953. Index numbers and indifference surfaces. *Trabajos de Estadística* 4: 209–241.
McFadden, D. 1978. Cost, revenue and profit functions. Chapter 1.1, 3–109, in Fuss and McFadden (1978).
McKenzie, L.W. 1957. Demand theory without a utility index. *Review of Economic Studies* 24: 185–189.
McKenzie, L.W. 1981. The classical theorems on existence of competitive equilibrium. *Econometrica* 49: 819–841.
Minkowski, H. 1911. Theorie der konvexen Körper. In *Gesammelte Abhandlungen*, 131–229. Leipzig/Berlin: Teubner II.
Moreau, J.-J. 1967. Fonctionelles convexes. Séminaire sur les equations aux derivées partielles, II. College de France, Mimeo.
Phelps, R.R. 1963. Support cones and their generalizations. In *Convexity*, Proceedings of Symposia in Pure Mathematics, vol. VII, ed. V.L. Klee. Providence: American Mathematical Society.
Rädström, H. 1949–50. II. Polar reciprocity. Seminar on Convex Sets. Princeton: Institute for Advanced Study, mimeo.
Robertson, A.P., and W. Robertson. 1964. *Topological vector spaces*, Cambridge Tracts in Mathematics and Mathematical Physics, No. 53. Cambridge: Cambridge University Press.
Rockafellar, R.T. 1970. *Convex analysis*. Princeton: Princeton University Press.
Ruys, P.H.M. 1972. On the existence of an equilibrium for an economy with public goods only. *Zeitschrift für Nationalökonomie* 32: 189–202.
Ruys, P.H.M., and H.N. Weddepohl. 1979. Economic theory and duality. In *Convex analysis and mathematical economics*, Lecture Notes in Economics and Mathematical Systems. No. 168, ed. J. Kriens, 1–72. New York: Springer.
Shephard, R.W. 1953. *Cost and production functions*. Princeton: Princeton University Press.
Shephard, R.W. 1970. *Theory of cost and production functions*. Princeton: Princeton University Press.
Uzawa, H. 1964. Duality principles in the theory of cost and production. *International Economic Review* 5: 216–220.
Weddepohl, H.N. 1972. Duality and equilibrium. *Zeitschrift für Nationalökonomie* 32: 163–187.
Wold, H. 1943. A synthesis of pure demand analysis: Part II. *Skandinavisk Aktuarietidskrift* 26: 220–263.
Young, L.C. 1939. On an inequality of Marcel Riesz. *Annals of Mathematics* 40: 567–574.
Young, L.C. 1969. *Lectures on the calculus of variations and optimal control theory.* Philadelphia: W.B. Saunders Company.

Gayer, Arthur David (1903–1951)

Anna J. Schwartz

Arthur (Archie) Gayer was a rising star in economic research in the United States during the 1930s but his promise was not sustained in later years of his foreshortened life. Born of English parentage at Poona in British India on 19 March 1903, he was educated at St Paul's School, London, 1916–1921, and Lincoln College, Oxford (BA, 1925; D. Phil., 1930). On arriving in the United States in 1927 as a Rockefeller Foundation fellow, he did graduate work at Columbia University, where he taught economics at Barnard College from 1931 to 1940. From 1940 until his death on 17 November 1951 as the result of an automobile accident, he taught at Queens College of the City University of New York.

Gayer's research covered three main subjects: public works in the United States; monetary policy and economic stabilization; and United Kingdom growth and cyclical experience, 1790–1850. He had begun work in 1929 as an assistant to Leo Wolman at the National Bureau of Economic Research on a statistical analysis of the volume, distribution, and fluctuations of local and federal construction expenditures. Gayer's study, covering the period 1919–1934, was generally regarded as a thorough examination of the empirical record as well as a careful summary of the theory of public works as a stabilizing device. In his study of monetary policy, in which he reviewed the traditional and interwar gold standard, Gayer advocated British-US cooperation in establishing a satisfactory international monetary system over the dollar-sterling area. In honour of Irving Fisher's seventieth birthday, he edited a volume of essays on the lessons of monetary experience. From 1936 to 1941, Gayer directed a study designed to expand his doctoral thesis, 'Industrial Fluctuation and Unemployment in England, 1815–1850'. The product of this research, in which he collaborated with a team of young scholars, was published in two volumes after Gayer's death. Conceptually, the study was heavily influenced by ideas in the *General Theory*, although the cyclical analysis was based on measures pioneered by the National Bureau. By the time the study appeared, the National Bureau's method of analysis commanded attention neither in the United States nor in Britain.

Selected Works

1935a. *Public works in prosperity and depression*. New York: National Bureau of Economic Research.

1935b. *Economic stabilization: A study of the gold standard*. London: A. & C. Black.

1937. *The lessons of monetary experience: Essays in honor of irving fisher*. New York: Farrar & Rinehart.

1953. (With W.W. Rostow and Anna J. Schwartz). *The growth and fluctuation of the British economy, 1790–1850*, 2 vols. Oxford: Clarendon Press. Reprinted, Hassocks: Harvester. 1975.

Gearing

J. S. S. Edwards

The question whether a company's choice of the proportion of debt to equity finance in its capital structure matters has involved a great deal of controversy. This choice, known as the gearing decision in the UK and the leverage decision in the USA, is widely regarded by corporate finance directors, investors, stock market participants and many others as an issue of considerable importance, yet the basic result of conventional economic theory applied to this question is that the gearing decision is irrelevant - there is no advantage to a firm in choosing one debt-equity ratio rather than another. This striking contrast between theory and practice has, of course, led to much critical examination of the assumptions of the theory, and some progress has been made in identifying ways in which gearing may matter. However it remains true that the determinants of a firm's gearing decision, and its importance, are not yet fully understood.

The argument that the gearing decision is a matter of irrelevance, affecting neither the firm's value nor its cost of capital (and hence its investment decision), is due to Modigliani and Miller in a celebrated article (Modigliani and Miller 1958). Their fundamental insight was that, in a world of perfect and complete capital markets in which taxation and asymmetric information are absent, individual investors can create any particular pattern of returns from holdings of securities by borrowing on their own account. This ability of investors to engage in 'home-made leverage' means that there is no reason for firms to concern themselves about the amount of debt in their capital structure: investors can create for themselves any pattern of returns which would be given by a share in a firm with a particular gearing ratio, so firms cannot gain by offering one such pattern rather than another.

To see this, consider the following simple illustration of the Modigliani-Miller argument (based

on Nickell 1978). Suppose that a firm possesses assets which yield 1000 $\widetilde{\theta}$ per annum in perprtuity, where $\widetilde{\theta}$ is a random variable, and that this firm has 1000 equity shares outstanding, but no debt in its capital structure. The price of a claim to an income stream yielding $\widetilde{\theta}$ in perpetuity is determined by a perfect capital market to be 1, so the value of the firm's equity is 1000. If the firm borrows, say, 200 at a rate of 10% per annum with no risk of default each share will now yield $\widetilde{\theta} - 0.02$ per annum in perpetuity, because there is a certain interest payment to be made from the returns on the firm's assets in each year before shareholders receive anything. If individual investors can borrow at the same interest rate as the firm the price of a share will now be 0.8, since an investor could have created a $\widetilde{\theta} - 0.02$ income stream in the original situation by borrowing 0.2 (on which the annual interest payment is 0.02) and using 0.8 of his own to buy one share for 1. Thus if the firm does borrow 200 the value of its equity will fall to 800, and its overall value (the sum of the values of outstanding debt and equity) remains constant at 1000. Home-made leverage enables an individual investor to create any combination of $\widetilde{\theta}$ and a certain return whatever combination of $\widetilde{\theta}$ and a certain return the firm offers, so that there is no reason for the firm to concern itself with the choice of a particular combination of the two.

Modigliani-Miller's original argument rested on a number of restrictive assumptions, such as the existence of risk classes within which firm's operating earnings were perfectly correlated, which were relaxed in subsequent work. One of the most general proofs of the Modigliani-Miller theorem was that by Stiglitz (1974), which did not need to make any assumptions about the existence of risk classes, the source of uncertainty, individuals having the same expectations, or the interest rate paid by a firm being unaffected by the amount of capital it raises. Stiglitz noted three critical limitations to his proof however, and it is these which begin to suggest ways in which gearing may be a relevant decision for firms. One is that individuals' expectations about future prices and firm valuations must not be affected by changes in companies' financial policy: this in effect rules out the possibility of financial policy acting as a signal in a world of asymmetric information. The second limitation is that individual borrowing must be a perfect substitute for firm borrowing, while the third is that there must be no bankruptcy.

At first sight these last two limitations would appear to indicate clearly why gearing is important in practice. But matters are not so straightforward. Companies may be able to borrow on better terms than individuals, but this may be because they are better risks: the Modigliani-Miller theorem only requires that individuals and firms borrow on the same terms for debt of equivalent risk. It is certainly not true that companies can always borrow on better terms than individuals: mortgages for house purchase, for example, are sometimes available at rates below those charged on corporate borrowing. Even if it is true that firms can borrow on better terms than individuals for equivalent risk loans, so that they can gain in value by offering this service to individuals, firms can compete by so doing and this may eliminate the gain in value: the supply of corporate debt expands until the Modigliani-Miller proposition is re-established.

A similar argument applies in the case of bankruptcy. When a firm issues risky debt it creates a security which individuals, lacking limited liability, cannot replicate by borrowing on their own account. This expands the range of portfolio opportunities open to investors, which they should in principle be willing to pay for thus enabling the firm to increase its value by use of some debt finance. However the extent to which a particular firm can gain by issuing risky debt depends on whether it can offer something special to investors that is not already available; it is difficult to believe that one more firm's risky debt significantly expands the set of portfolio opportunities available to investors.

A more promising approach to understanding the importance attached to the gearing decision in practice would appear to be the relaxation of the assumption that there is no taxation. Most corporation tax systems allow interest payments on debt finance to be deductible against corporation tax, and this has been widely argued to provide the obvious explanation for the use of some debt in a

firm's capital structure (by Modigliani-Miller among many others). Tax advantages to debt might imply that firms should use all-debt finance, but this unsatisfactory conclusion has been avoided by introducing costs of bankruptcy and financial distress, which reduce the size of the total payout to investors in certain contingencies that are more likely the larger the firm's gearing ratio. These include costs of reorganization and liquidation associated with bankruptcy, together with costs of financial distress, such as the foregoing of profitable investment opportunities which may be necessitated by bankruptcy, and the making of suboptimal investment decisions in an attempt to forestall bankruptcy. These costs will result in the firm's market value beginning to decline beyond some level of gearing, so that together with the corporation tax advantage of debt a theory of optimal gearing ratios seems to result.

This theory is perhaps the most commonly accepted one for explaining the importance of gearing, but it is by no means uncontroversial. One reason for its lack of universal acceptance is that evidence on the size of the costs of bankruptcy and financial distress is limited and, where available (Warner 1977), it does not suggest that they are large. Another reason is that this theory only takes account of corporate taxes in arguing that there is a tax advantage to the use of debt finance. Investors are subject to personal taxes on interest and dividend income and capital gains, and these tax rates differ usually between income and capital gains, and certainly between different investors. This causes a number of problems. If personal tax is higher on debt interest than on equity income (taking account of both dividends and capital gains) the use of debt finance may reduce the firm's value. The variation of personal tax rates across investors means that it is likely that some would prefer debt on tax grounds while others would prefer equity. Indeed differences in personal tax rates were used by Miller (1977) to reintroduce an irrelevance result (in which bankruptcy costs were ignored): he argued that investors would specialize their holdings in debt or equity according to whether their after-all-tax income from a unit of pre-tax debt cash flow (1 – personal tax rate on debt income) was greater or less than their after-tax income from a unit of pre-tax equity cash flow ((1 – personal tax rate on equity income) times (1-corporation tax rate)). There would be a determinate aggregate debt-equity ratio, at which point marginal investors would be indifferent between holding debt or equity on tax grounds, but the gearing decision would be irrelevant for individual firms.

Miller's argument shows that when heterogeneous personal tax rates are considered, as they must be, it is not obvious that there is a tax advantage to corporate borrowing. But there are problems with this argument too. Auerbach and King (1983) show that the Miller equilibrium requires the existence of certain constraints on investors: without such constraints (on, for example, borrowing and short-selling) questions arise concerning the existence of an equilibrium, for with perfect capital markets realistic tax systems provide opportunities for unlimited arbitrage at government expense between investors and firms in different tax positions. Auerbach and King also show that the combined effect of taxation and risk is to produce a situation in which gearing is relevant. With individual investors facing different tax rates and wishing to hold diversified portfolios the Miller equilibrium can no longer be sustained: investors who on tax grounds alone would hold only equity may nevertheless hold some debt because an equity-only portfolio would be too risky. Miller's argument also does not take account of the implications of uncertainty and the asymmetric treatment of profits and losses which is a feature of most corporation tax systems. De Angelo and Masulis (1980) argued that the probability of interest tax shields being lost or deferred precluded the Miller equilibrium, and suggested that an optimal gearing ratio existed for a firm where the cost of debt finance, taking account of the probability of being unable to offset interest fully against corporation tax, equalled the cost of equity.

Miller's argument should thus be seen as one which raises important questions about whether

taxation really does give incentives for individual firms to use debt finance, but does not clearly establish that there are no such incentives. It therefore weakens the theory based on trading off tax advantages of debt against costs of bankruptcy and financial distress, but does not destroy it. Another way in which this theory has been weakened is as a result of work on capital structure and financial policy which drops the assumption that the probability distribution of a firm's profits is common knowledge and independent of the firm's financial structure (this is essentially the first of the three critical limitations to Stiglitz' proof of the Modigliani-Miller theorem discussed above). There are a number of models based on asymmetric information of one sort or another in which a firm's gearing decision is not irrelevant. One type of model is where the firm's managers know more about the firm's possible returns than do outside investors. Ross (1977) assumes that managerial rewards depend on the current value of the firm and its future returns, and managers know the distribution of future returns while outside investors do not. The amount of debt chosen acts as a signal: managers of firms with higher expected future returns choose larger amounts of debt because only the managers of the better firms are willing to incur the increased risk of bankruptcy and its related costs associated with higher debt. Another type of model is based on principal-agent considerations: firms are run by managers (agents) on the behalf of shareholders (principals), but managers have some scope for pursuing their own interests at shareholders' expense because of asymmetric information. This is, however, recognized by the shareholders. The general form of models of this type is that managers choose a financial structure of the firm which determines managerial incentives: the capital market understands the incentives implied by a particular financial structure, and values the firm accordingly: this evaluation is taken into account by the managers in choosing financial structure. It is clear how a determinate capital structure can emerge from this framework, and Jensen and Meckling (1976) and Grossman and Hart (1982) are two examples of papers where gearing is important because of these reasons.

This work on asymmetric information and capital structure is highly suggestive of factors which may make gearing important, but as yet all we have in this area are insights rather than a complete and coherent theory. In particular there has been little integration of the traditional taxation arguments into the asymmetric information approach. Hence economists' understanding of firms' gearing decisions is still imperfect.

See Also

- ▶ Dividend Policy
- ▶ Finance
- ▶ Retention Ratio

Bibliography

Auerbach, A.J., and M.A. King. 1983. Taxation, portfolio choice and debt-equity ratios: A general equilibrium model. *Quarterly Journal of Economics* 98(4): 587–609.

De Angelo, H., and R.W. Masulis. 1980. Optimal capital structure under corporate and personal taxation. *Journal of Financial Economics* 8(1): 3–29.

Grossman, S.J., and O.D. Hart. 1982. Corporate financial structure and managerial incentives. In *Economics of information and uncertainty*, ed. J. McCall. Chicago: University of Chicago Press.

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.

Miller, M.H. 1977. Debt and taxes. *Journal of Finance* 32(2): 261–275.

Modigliani, F., and M.H. Miller. 1958. The cost of capital, corporation finance, and the theory of investment. *American Economic Review* 48: 261–297.

Nickell, S.J. 1978. *The investment decisions of firms*. Cambridge: Cambridge University Press.

Ross, S.A. 1977. The determination of financial structure: The incentive-signalling approach. *Bell Journal of Economics* 8(1): 23–40.

Stiglitz, J.E. 1974. On the irrelevance of corporate financial policy. *American Economic Review* 64(6): 851–866.

Warner, J.B. 1977. Bankruptcy costs: Some evidence. *Journal of Finance* 32(2): 337–347.

Geary, Robert Charles (1896–1983)

J. R. N. Stone

Keywords

Geary, R. C.; Instrumental variables; Ireland, economics in; Linear expenditure system; Methodology of econometrics

JEL Classifications

B31

Geary was born on 11 April 1896 in Dublin, and died on 8 February 1983, also in Dublin. He was educated at University College, Dublin, from 1913 to 1918, and at the Sorbonne from 1919 to 1921. In 1922 he became assistant lecturer in mathematics at University College, Southampton. He held the post of statistician in the Department of Industry and Commerce in Dublin from 1923 to 1949. From 1946 to 1947 he was a Senior Research Fellow in the Department of Applied Economics in Cambridge. He held the position of First Director in the Central Statistical Office of Eire from 1949 to 1957. In 1957 he was appointed Head of the National Accounts Branch of the United Nations Statistical Office, which post he held until 1960. From 1960 to 1966 he was director of the newly founded Economic (and Social) Research Institute in Dublin, remaining attached to it as consultant from 1966 until his death.

Geary was a mathematical statistician of international standing. His statistical writings cover a wide range of topics, among them testing for normality, the distribution of ratios, parameter estimation, and so on, some of which are relevant to econometric methodology. Indeed, much of his work has an explicitly economic content: he was probably responsible for the excellent report prepared by the Department of Industry and Commerce containing the first national accounts of Eire (Eire, Minister for Finance, 1946); he wrote many papers on the determination of relationships between variables, notably Geary (1948, 1949), in the second of which he uses instrumental variables; he derived the form of the utility function underlying the linear expenditure system (Geary 1950–51); he was part-author of a monograph on linear programming applied to economics (Geary and McCarthy 1964); and he built a model of the Irish economy based on an accounting framework (Geary 1963–4). His published output numbers 112 titles, more than half of which appeared after his 65th birthday. A full bibliography is appended to Spencer (1976, 1983).

Selected Works

1946. Eire, Minister for Finance. *National income and expenditure* 1938–1944. Dublin: Stationery Office.

1948. Studies in relations between economic time series. *Journal of the Royal Statistical Society, Series B* 10(1): 140–158.

1949. Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown. *Econometrica* 17: 30–58.

1950–51. A note on 'A constant utility index of the cost of living'. *Review of Economic Studies* 18: 65–66.

1964. (With M.D. McCarthy.) *Elements of linear programming with economic applications*, 2nd ed. London: Griffin. (with J.E. Spencer), 1973.

1963–4. Towards an input–output decision model for Ireland. *Journal of the Statistical and Social Inquiry Society of Ireland* 21(Pt 2): 67–119.

Bibliography

Spencer, J.E. 1976. The scientific work of Robert Charles Geary. *Economic and Social Review* 7 (3): 233–247.

Spencer, J.E. 1983. Robert Charles Geary – An appreciation. *Economic and Social Review* 14 (3): 161–164.

Gee, Joshua (*fl.*1725–1750)

Douglas Vickers

Keywords

Child, J.; Consumers' expenditure; Employment of the poor; Gee, J.; Labour supply; Protectionism; Workhouses

JEL Classifications

B31

Joshua Gee's place in the history of economics rests on his contributions to the protectionist literature during the early decades of the 18th century. He collaborated with Henry Martin among others in publishing the *British Merchant* that argued the protectionist case in 1713 and 1714 against the Treaty of Commerce proposed at Utrecht, and he published an extensive discussion of England's foreign trade, together with strong protectionist sentiments, in *The Trade and Navigation of Great Britain Considered* in 1729.

Addressing the current decline in the English export trades, the high level of imports of certain commodities, the demand for which, particularly in the case of French fashion goods, could be met by home produced import substitutes, the declining health of the woollen industry, and the currently widespread unemployment, Gee made a number of proposals for government regulation of trade and manufacturing. These proposals were directed principally to the need for 'finding effectual ways for employing the poor', thereby aligning his work with the widespread employment argument of the time. To the same end he advocated also a wider development of workhouses. Following Josiah Child, Gee proposed that trade with the colonial plantations should be regulated in such a way as not only to encourage their production of the materials needed for English manufacturing industries, thereby 'employing all the poor', but also to facilitate 'supplying our plantations with everything they want and all manufactured within ourselves'.

Gee's descriptive essays exhibited an understanding of the interdependence of economic activities and processes, 'one employment depending on another', and of 'the circulation of commerce that must infuse riches into every part'. He argued that higher domestic commodity prices would induce workers to increase their supply of labour, leading to higher incomes and also higher discretionary consumption expenditures. But such potentially important notions were not accorded any systematic or analytical development.

See Also

► Mercantilism

Selected Works

1729. *The trade and navigation of Great Britain considered.* London.

1742. *An impartial enquiry into the importance and present state of the woollen manufactures of Great Britain.* London.

1742. *The Grazier's advocate, or free thoughts of wool and the woollen trade.* London.

Gender

Francine D. Blau

The term gender has traditionally referred, as has sex, to the biological differences between men and women. More recently a movement has arisen both in social science writings and in public discourse to expand this definition to encompass also the distinctions which society has erected on this biological base, and further to use the word gender in preference to sex to refer to this broader definition. In this essay, we describe the relationship of

this expanded concept of gender to economic theory.

Historically, gender has not been perceived to be a central concept in economic analysis, either among the classical and neoclassical schools or among Marxist economists. However, as the force of current events has thrust gender-related issues to the force, economists have responded by seeking to analyse these issues. The outcome of this process has been not only a better understanding of the nature of gender differences in economic behaviour and outcomes, but also an enrichment of the discipline itself.

While, as noted above, the mainstream of economic analysis paid scant attention to genderrelated issues, the 19th century campaign for female suffrage did focus some attention on gender inequality. Among classical economists, J.S. Mill (1878) eloquently argued for the 'principle of perfect equality' (p. 91) between men and women. Not only did he favour equality of the sexes within the family, but also women's 'admissibility to all the functions and occupations hitherto retained as the monopoly of the stronger sex'. He also expressed the belief 'that their disabilities elsewhere are only clung to in order to maintain their subordination in domestic life' (p. 94). In the Marxist school, Engels (1884) tied the subjection of women to the development of capitalism and argued that women's participation in wage labour outside the home, as well as the advent of socialism, was required for their liberation. The belief in the emancipating effects of a fuller participation in employments outside the home was shared not only by Mill and Engels, but also by such contemporary feminist writers as Gilman (1898).

The passage of time has proved these views oversimplified. As Engels and Gilman correctly foresaw, there has been an increase in the labour force participation of women, particularly of married women, in most of the advanced industrialized countries. This has undoubtedly altered both the relationship between men and women and the very organization of society in many ways. However, while women's labour force participation has in many instances risen dramatically, it nonetheless remains the case that the types of jobs held by men and women as well as the earnings they receive continue to differ markedly.

The contribution of modern neoclassical analysis, which comprises the main focus of this essay, has been to subject to greater scrutiny and more rigorous analysis both women's economic roles within the family and the causes of gender inequality in economic outcomes. We examine each of these areas below. However, the interrelationships between the family and the labour market, most importantly the consequences of labour market discrimination against women for their roles and status in the family, have tended to be neglected. Nonetheless, the possible existence of such feedback effects is an important issue which is also considered here.

Time Allocation in the Family Context

Prompted in part by their desire to understand the causes of the rising labour force participation of married women in the post-World War II period, economists extended the traditional theory of labour supply to consider household production more fully. The consequence was not only a better understanding of the labour supply decision, but also the development of economic analyses of the related phenomena of marriage, divorce and fertility.

The Traditional Theory of Labour Supply

The traditional theory of labour supply, also known as the labour-leisure dichotomy, was a simple extension of consumer theory. In this model, individuals maximize their utility, which is derived from market goods and leisure, subject to budget and time constraints. Where an interior solution exists, utility is maximized when the individual's marginal rate of substitution of income for leisure is set equal to the market wage.

Since in this model all time not spent in leisure is spent working, a labour supply (leisure demand) function may be derived with the wage, non-labour income, and tastes as its arguments. The well-known results of consumer theory are

readily obtained. An increase in non-labour income, all else equal, increases the demand for all normal goods including leisure, inducing the individual to consume more leisure and to work fewer hours (the income effect). An increase in the wage, ceteris paribus, has an ambiguous effect on work hours due to two opposing effects. On the one hand, the increase in the wage is like an increase in income and in this respect tends to lower work hours due to the income effect. On the other hand, the increase in the wage raises the price (opportunity cost) of leisure inducing the individual to want to consume less of it, i.e., a positive substitution effect on work hours.

The theory sheds light on the labour force participation decision when it is realized that a corner solution will arise if the marginal rate of substitution of income for leisure at zero work hours is greater than the market wage. In this case, the individual maximizes utility by remaining out of the labour force. The impact of an increase in the wage is unambiguously to raise the probability of labour force participation, since, at zero work hours, there is no off-setting income effect of a wage increase.

Household Production and the Allocation of Time

While the simple theory is sufficient for some purposes, it has limited usefulness for understanding the determinants of the gender division of labour in family and the factors influencing women's labour force participation, both at a point in time and trends over time. The key to addressing these issues is a fuller understanding and analysis of the household production process.

The first step in this direction was taken by Mincer (1962) who pointed out the importance, especially for women, of the three-way decision among market work, non-market work and leisure. He argued that the growth in married women's labour force participation was due to their rising real wages which increased the opportunity cost of time spent in non-market activities. But since, during the same period, the real wages of married men were also increasing, this must mean that the substitution effect associated with women's own real wage increases dominated the income effect associated with the growth in their husbands' real wages. While this part of the analysis could be accommodated in the framework of the traditional model, the next question Mincer raised could not. Why should the substitution effect dominate the income effect for women when such time series evidence as the declining work week suggested a dominance of the income effect over the substitution effect for men? The answer, according to Mincer, lay in women's responsibility for non-market production. The opportunities for substituting market time (through the purchase of market goods and services) are greater for time spent in home work than for time spent in leisure. Thus, since married women spend most of their non-market time on household production while men spend most of theirs on leisure, the substitution effect of a wage increase would be larger for married women than for men.

Becker (1965) advanced this process considerably by proposing a general theory of the allocation of time to replace the traditional theory of labour supply. In this and other work (summarized in Becker 1981), he laid the foundations of what has become known as the 'new home economics', and spearheaded the development of economic analyses of time allocation, marriage, divorce and fertility. Interestingly, while Mincer opened a window on household production by distinguishing non-market work from leisure where the traditional labour supply theory had not done so, Becker was able to provide a further advance by again eliminating the distinction. However, while in the traditional labour supply model all non-market time is spent in leisure, in Becker's model all non-market time is spent in household production.

Specifically, Becker assumes that households derive utility from 'commodities' which are in turn produced by inputs of market goods and non-market time. It is interesting to note that Becker's 'commodities', produced and consumed entirely in the home, are the polar opposite of Marx's (1867) 'commodities', produced and exchanged in the market. Examples of Becker's

commodities range from sleeping, which is produced with inputs of non-market time and of market goods like a bed, sheets, a pillow and a blanket (and in some cases, perhaps, a sleeping pill); to a tennis game that is produced by inputs of non-market time combined with tennis balls, a racquet, an appropriate costume, and court time; to a clean house produced with inputs of non-market time and a vacuum cleaner, a bucket and a mop, and various cleaning products.

In this model the production functions for the commodities are added to the constraints of the utility maximization problem. Utility can still be expressed as a function of the quantities of market goods and non-market time consumed; however, market goods and non-market time now produce utility only indirectly through their use in the production of commodities. Relative preferences for market goods versus home time depend on the ease with which the household can substitute market goods for non-market time in consumption and production. Substitution in consumption depends on their preferences for 'goods intensive' commodities – those produced using relatively large inputs of market goods in comparison to non-market time – relative to 'time intensive' commodities – those produced using relatively large inputs of non-market time in comparison to market goods. Substitution in production depends on the availability of more goods-intensive production techniques for producing the same commodity.

The usefulness of these ideas may be illustrated by considering the relationship of children to women's labour force participation. Children (especially when they are small) may be viewed as a time-intensive 'commodity'. Traditionally, it has been the mother who has been the primary care-giver. Moreover, while it is possible to substitute market goods and services for home time in caring for children (in the form of babysitters, day care centres, etc.), these alternative production techniques tend to be costly and it is sometimes difficult to make suitable alternative arrangements (in terms of quality, scheduling, etc.). Thus, at a point in time, the probability that a woman will participate in the labour force is expected to be inversely related to the number of small children present. Over time, the increase in women's participation rates has been associated with decreases in birthrates, as well as increases in the availability of various types of child care facilities, formal and informal. Changes in social norms (Brown 1984) making it more acceptable to substitute for the time of parents in the care of young children may also have been a factor, although it is difficult to know, in this case as in others, the extent to which attitude change precedes or follows change in the relevant behaviour.

The relationship between labour force participation and fertility is reinforced by the impact of the potential market wage on women's fertility decisions. Greater market opportunities for women have increased the opportunity cost of children (in terms of their mothers' time inputs) and induced families to have fewer of them. Similarly, the greater demand for alternative child care arrangements (also due to the increased value of women's market time) has made it profitable for more producers to enter this sector.

The Gender Division of Labour

In our discussion of children, we simply assumed that women tend to bear the primary responsibility for child care. However, the gender division of labour in the family is also an issue which the new home economics addresses. According to Becker (1981), the division of labour will be dictated by comparative advantage. To the extent that women have a comparative advantage in household and men in market production, it will be efficient for women to specialize to some extent in the former while men specialize in the latter. In this view, the increased output corresponding to this arrangement constitutes one of the primary benefits to marriage. Thus, women's increasing labour force participation is seen to have reduced the gains from marriage thereby contributing to the trend towards higher divorce and lower marriage rates. The notion that, where families are formed, it is generally efficient and thus optimal for one member, usually the wife, to specialize to some extent in household production, while the husband specializes entirely in market work, has important consequences for women's status in the labour

market. As we shall see in greater detail below, human capital theorists expect such a division of labour to lower the earnings of women relative to men, due to work force interruptions and smaller investments in market-oriented human capital. For this and other reasons, it is important to consider in greater detail whether such specialization is indeed as desirable for the family as the model suggests, and, by implication, whether it is apt to continue into the future. There are three points to be made in this regard.

First, such a division of labour may not be as advantageous for women as it is for men (Ferber and Birnbaum 1977; Blau and Ferber 1986). Thus, even if such a specialization is efficient in many respects, it may not maximize the family's utility. Indeed, when there are conflicts of interest or even pronounced differences in tastes between the husband and wife, the concept of the family utility function itself becomes less meaningful, since the way in which the preferences of family members can meaningfully be aggregated to form such a utility function has not been satisfactorily specified.

What are the disadvantages to women of their partial specialization in household production? First, in a market economy, such an arrangement makes them to a greater or lesser degree economically dependent on their husbands (see also Hartmann 1976). This is likely to reduce their bargaining power relative to their husbands' in family decision-making, as well as to increase the negative economic consequences for them (and frequently for their children) of a marital break-up. In the face of recent increases in the divorce rate, such specialization has become a particularly risky undertaking. Second, as more women come to value their careers in much the same way as men do, both in terms of achievement and earnings, their specialization in housework to the point where it is detrimental to their labour market success is not apt to be viewed with favour by them. The utility-maximizing family will take these disadvantages into account in conjunction with the efficiency gains of specialization in allocating the time of family members.

If specialization is indeed considerably more productive than sharing of household responsibilities, it may be possible for this higher output to be used in part to compensate women for the disadvantages detailed above. However, it is likely that the gains to such specialization will shrink over time relative to the disadvantages of such an arrangement. As women anticipate spending increasingly more of their working lives in the labour market, their investments in market-oriented capital may be expected to continue to grow and their comparative advantage in home work relative to men to decline. Moreover, as the quality of the opportunities open to women in the labour market continues to improve, the disadvantages of specialization in home work in the form of foregone earnings and possibilities for career advancement will also rise. Thus, greater sharing of household responsibilities between men and women is likely to become increasingly prevalent, even if women in general retain a degree of comparative advantage in household production for some time to come.

A second point to be made with regard to women's specialization in household production is that comparative advantage does not comprise the only economic benefit to family or household formation (Ferber and Birnbaum 1977; Blau and Ferber 1986). Families and households also enjoy the benefits of economies of scale in the production of some commodities, as well as the gains associated with the joint consumption of 'public' goods. These benefits of collaboration would be unaffected by a reduction in specialization, even if those based on comparative advantage would be diminished. Other benefits of marriage or household formation may actually be increased by a more egalitarian division of household responsibilities. For example, two-earner families are in a sense more diversified and thus enjoy greater income security than families which depend on only one income. It may also be the case that the enjoyment derived from joint consumption is enhanced when the members of a couple have more in common, as when both participate in market and home activities. Thus, the incentives of couples to adhere to the traditional division of labour in order to enjoy the economic benefits of marriage may not be as strong as suggested

when only the gains to comparative advantage are considered.

Finally, it is important to point out that women's comparative advantage for household production may stem not only from the impacts of biology and gender differences in upbringing and tastes, but also from the effect of labour market discrimination in lowering women's earnings relative to men's. Decisions based to some extent on such market distortions are not optimal from the perspective of social welfare even though they may be rational from the perspective of the family. The importance of such feedback effects are considered in greater detail below.

Gender Differences in Labour Market Outcomes

We turn now to the contribution of economic analysis to an understanding of the causes of gender inequality in economic outcomes. Here, the consideration of gender issues has been accommodated principally through the development of new and interesting applications of existing theoretical approaches. The particular challenge posed to the theories by women's economic status is the existence of occupational segregation as well as earnings differentials by sex. Occupational segregation refers to the concentration of women in one set of predominantly female jobs and of men in another set of predominantly male jobs. The reasons for such segregation and its relationship to the male–female pay differential are two key questions to be addressed.

As in the case of the analysis of women's roles in the family, the catalyst for the development of these approaches was provided by external events. Some moderate degree of interest in this issue was generated in England by the World War I experience. Pursuant to the war effort, there was some substitution of women into traditionally male civilian jobs, although not nearly to the degree that there would be during World War II. Questions of the appropriate pay for women under these circumstances arose and stimulated some economic analyses of the gender pay differential – all of which gave a prominent causal role to occupational segregation. These included the work of Fawcett (1918) and Edgeworth (1922) (which provided the antecedents for Bergmann's (1974) overcrowding model, discussed below) and Webb (1919). The analysis of gender differentials in the labour market received another impetus in the early 1960s, this time in the United States, with the development of the women's liberation movement and the passage of equal employment opportunity legislation. Two broad approaches to the issue have since evolved. First is the human capital view which lays primary emphasis on women's own voluntary choices in explaining occupation and pay differences. Second are a variety of models of labour market discrimination which share the common characteristic of placing the onus for the unequal outcomes on differential treatment of equally (or potentially equally) qualified men and women in the labour market. While these two approaches may be viewed as alternatives, it is important to point out that they are in fact not mutually exclusive. Both may play a part in explaining sex differences in earnings and occupations and the empirical evidence suggests that this is the case (see, e.g., Treiman and Hartmann 1981). Indeed, as we shall see, their effects are quite likely to reinforce each other. We now consider each of these approaches in turn.

The Human Capital Explanation

The human capital explanation for gender differences in occupations and earnings, developed by Mincer and Polachek (1974), Polachek (1981) and others, follows directly from the analysis of the family described above. It is assumed that the division of labour in the family will result in women placing greater emphasis that men on family responsibilities over their life cycle. Anticipating shorter and more discontinuous work lives as a consequence of this, women will have lower incentives to invest in market-oriented formal education and on-the-job training than men. Their resulting smaller human capital investments will lower their earnings relative to those of men.

These considerations are also expected to produce gender differences in occupational distribution. It is argued that women will choose occupations for which such investments are less important and in which the wage penalties associated with work force interruptions (due to the skill depreciation that occurs during time spent out of the labour force) are minimized. Due to their expected discontinuity of employment, women will avoid especially those jobs requiring large investments in firm-specific skills (i.e. skills which are unique to a particular enterprise), because the returns to such investments are reaped only as long as one remains with the firm. The shorter expected job tenure of women in comparison with that of men is also expected to make employers reluctant to hire women for such jobs in that employers bear some of the costs of such training. Thus, to the extent that it is difficult to distinguish more from less career-oriented women, the former may be negatively affected (see the discussion of statistical discrimination below).

More recently, Becker (1985) has further argued that, even when men and women spend the same amount of time on market jobs, women's homemaking responsibilities can still adversely affect their earnings and occupations. Specifically, he reasons that since child care and housework are more effort intensive than are leisure and other household activities, married women will spend less effort than married men on each hour of market work. The result will be lower hourly earnings for married women and, to the extent that they seek less demanding jobs, gender differences in occupations.

Thus, the human capital analysis provides a logically consistent explanation for gender differences in market outcomes on the basis of the traditional division of labour by gender in the family. An implication generally not noted by those who have developed this approach is that, to the extent that the human capital explanation is an accurate description of reality, it serves to illustrate graphically the disadvantages for women of responsibility for (specialization in) housework which we discussed above. To the extent that gender differences in economic rewards are not fully explained by productivity differences, we must turn to models of labour market discrimination to explain the remainder of the difference.

Models of Labour Market Discrimination

As noted earlier, models of discrimination were developed to understand better the consequences of differences in the labour market treatment of two groups for their relative economic success. The starting point for models of labour market discrimination is the assumption that members of the two groups are equally or potentially equally productive. That is, except for any direct effects of the discrimination itself, male and female labour (in this case) are perfect substitutes in production. This assumption is made not because it is necessarily considered an accurate description of reality, but rather because of the question which discrimination models specifically address: why do equally qualified male and female workers receive unequal rewards? Such models may then be used to explain how discrimination can produce pay differentials between men and women in excess of what could be expected on the basis of productivity differences.

Theoretical work in this area was initiated by Becker's (1957) model of racial discrimination. Becker conceptualized discrimination as a taste or personal prejudice. He analysed three cases, those in which the tastes for discrimination were located in employers, co-workers and customers, respectively. As Becker pointed out, for such tastes to affect the economic status of a particular group adversely, they must actually affect the behaviour of the discriminators.

One may at first question whether such a model is as applicable to sex as to race discrimination in that, unlike the case of racial discrimination, men and women are generally in close contact within families. However, the notion of socially appropriate roles, not explicitly considered by Becker, both sheds light on this question and establishes a link between his theory and occupational segregation. Thus, employers may be quite willing to hire women as secretaries, receptionists or nursery school teachers but may be reluctant to employ

them as lawyers, college professors or electricians. Co-workers may be quite comfortable working with women as subordinates or in complementary positions, but feel it is demeaning or inappropriate to have women as supervisors or as peers. Customers may be happy to have female waitresses at a coffee shop, but expect to be served by male waiters at an elegant restaurant. They may be delighted to purchase women's blouses or even men's ties from female clerks, but prefer their appliance salesperson, lawyer or doctor to be a man. Such notions of socially appropriate roles are quite likely a factor in racial discrimination as well.

Employers with tastes for discrimination against women in particular jobs will be utility rather than profit maximizers. They will see the full costs of employing a woman to include not only her wages but also a discrimination coefficient ($d_r \geq 0$) reflecting the pecuniary value of the disutility caused them by her presence. Thus, they will be willing to hire women only at lower wages than men ($w_f = w_m - d_r$). If men are paid their marginal products, employer discrimination will result in women receiving less than theirs. When employers differ in their tastes for discrimination, the market-wide discrimination coefficient will be established at a level which equates supply and demand for female labour at the going wage. Thus the size of the male–female pay gap will depend on the number of women seeking work, as well as on the number of discriminatory employers and on the magnitude of their discrimination coefficients.

One of the particularly interesting insights of Becker's (1957) analysis is that profit-maximizing employers who do not themselves have tastes for discrimination against women will nonetheless discriminate against them if their employees or customers have such prejudices. Male employees with tastes for discrimination against women will act as though their wage is reduced by $de(\geq 0)$, their discrimination coefficient, when they are required to work with women. Thus, they will consent to be employed with women only if they receive a higher wage – in effect a compensating wage differential for this unpleasant working condition.

The obvious solution to this problem from the employer's point of view is to hire a singlesex work force. If all employers followed such a strategy, male and female workers would be segregated by firm, but there would be no pay differential. Yet, as Arrow (1973) has noted, employers who have made a personnel investment in their male workers, in the form of recruiting, hiring or training costs, may not find it profitable to discharge all their male employees and replace them with women, even if the latter become available at a lower wage. While such considerations cannot explain how occupations initially become predominantly male, it can shed light on one factor – the necessity of paying a premium to discriminatory male workers to induce them to work with women – contributing to the perpetuation of that situation. Further, where women do work with discriminating male workers, a pay differential will result.

Some extensions of Becker's (1957) analysis of employee discrimination are also of interest. Bergmann and Darity (1981) point out that employers may be reluctant to hire women into traditionally male jobs because of adverse effects on the morale and productivity of the existing male work force. Given the replacement costs discussed above this would be an important consideration. As Blau and Ferber (1986) note, employee discrimination may also directly lower women's productivity relative to that of men. For example, since much on-the-job training is informal, if male supervisors or coworkers refuse or simply neglect to instruct female workers in these job skills, women will be less productive than men workers. Similarly, the exclusion of women from informal networks and mentor-protégé relationships in traditionally male occupations can diminish their access to training experiences and even to the information flows needed to do their jobs well.

Customer discrimination can also reduce the productivity of female relative to male employees. Customers with tastes for discrimination against women will act as if the price of a good or service provided by a woman were increased by their discrimination coefficient, dc (≥ 0). Thus, at any given selling price, a female employee will bring in less revenue than a male employee.

Women either will not be hired for such jobs or will be paid less. The potential applicability of this model is not only to conventional sales jobs. In our 'service economy', a large and growing number of jobs entail personal contact between workers and customers/clients.

Models based on the notion of tastes for discrimination are consistent with occupational segregation, but do not necessarily predict it. If wages are flexible, it is altogether possible that such discrimination will result in lower pay for women, but little or no segregation. However, if discriminatory tastes against women in traditionally male pursuits (on the part of employers, employees and/or customers) are both strong and prevalent, women may tend to be excluded from these areas. On the other hand, even if such segregation occurs, it may or may not be associated with gender pay differentials. In the presence of sufficient employment opportunities in the female sector, equally qualified women may earn no less than men.

The relationship between occupational segregation and earnings differentials is further clarified in Bergmann's (1974) overcrowding model. If for whatever reason – labour market discrimination or their own choices – potentially equally qualified men and women are segregated by occupation, the wages in male and female jobs will be determined by the supply and demand for labour in each sector. Workers in male jobs will enjoy a relative wage advantage if the supply of labour is more abundant relative to demand for female than for male occupations. Such 'crowding' of female occupations can also widen differentials between male and female jobs that would exist in any case due to women's smaller human capital investments or to employers' reluctance to invest in their human capital.

Perhaps the most serious question that has been raised about the Becker analysis, particularly of the case of employer discrimination, is its inability to explain the persistence of discrimination in the long run. Assuming that tastes for discrimination vary, the least discriminatory firms would employ the highest proportion of lower-priced female labour. They would thus have lower costs of production and, under constant returns to scale, could in the long run expand and drive the more discriminatory firms out of business (Arrow 1973).

This issue has provided the rationale, at least in part, for the elaboration of alternative models of discrimination, including the statistical discrimination model discussed below. Others, not considered here, have emphasized non-competitive aspects of labour markets (e.g. Madden 1973). However, this criticism of the Becker model is a double-edged sword in that it has led some economists to doubt that labour market discrimination is responsible, in whole or part, for gender inequality in economic rewards. Yet it is important to recognize that the phenomenon which we seek to understand is intrinsically complex. From this perspective it is not surprising that no easy solution has been found to the question of why discrimination has persisted. Similarly, the various models of discrimination, each emphasizing different motivations and different sources of this behaviour, need not be viewed as alternatives. Rather, each may serve to illuminate different aspects of this complex reality.

As noted above, models of statistical discrimination were developed by Phelps (1972) and others to shed light on the persistence of discrimination. They do so by imputing a motive for employer discrimination which, in an environment of imperfect information, is consistent with profit maximization. Statistical discrimination occurs when employers believe that, all else equal, women are on average less productive or less stable workers than men. The common perception that women are more likely to quit their jobs than men would be an example of this.

As in the employer taste for discrimination model, statistical discrimination would cause employers to prefer male workers and to be willing to hire women only at a wage discount. A difference is, however, that in this case male and female workers are not perceived to be perfect substitutes. Further, if women are viewed as less stable workers, there will tend to be substantive differences between male and female jobs, with the former emphasizing firm-specific skills to a greater extent. This is essentially the picture painted by the dual market model (Piore 1971; Doeringer and Piore 1971). In this view, women

tend to be excluded from the 'primary sector', jobs requiring firm-specific skills and thus characterized by relatively high wages, good promotion opportunities and low turnover rates, and to find employment in the 'secondary sector', comprised of low paying, dead-end jobs in which there tends to be considerable turnover.

Like the human capital model, the notion of statistical discrimination provides a link between women's roles in the family and gender differences in market outcomes. However, the connection is in terms of differences in the treatment of men and women, rather than differences in the choices they make.

One crucial issue is of course whether employers' perceptions are indeed correct. If they are, as Aigner and Cain (1977) have pointed out, then in some sense labour market discrimination as conventionally defined does not exist: women's lower wages are due to their lower productivity. Nonetheless, the employer's inability to distinguish between more and less career-oriented women certainly creates an inequity for the former vis-à-vis their male counterparts.

On the other hand, employer perceptions may be incorrect or exaggerated. Differentials based on such erroneous views undoubtedly constitute discrimination as economists have defined it. However, as Aigner and Cain (1977) have persuasively argued, gender differentials based on employers' mistaken beliefs are even less likely to persist in the long run than those based on employers' tastes for discrimination. Nonetheless, in times of rapid changes in gender roles, there may be considerable lags in employers' perceptions. Employers' incorrect views could also magnify the impact of employee or customer discrimination, as when such discrimination is either less extensive or more susceptible to change than employers believe.

A potentially more powerful role for statistical discrimination is provided in models which allow for feedback effects, for example Arrow's (1973) model of perceptual equilibrium. In this case, men and women are assumed to be potentially perfect substitutes in production, but employers believe that, for example, women are less stable workers (Arrow 1976). They thus allocate women to jobs where the cost of turnover is minimized and women respond by exhibiting the unstable behaviour employers expect. The employers' assessments are correct *ex post*, but are in fact due to their own discriminatory actions. This equilibrium will be stable even though an alternative equilibrium is potentially available in which women are hired for jobs which are sufficiently rewarding to inhibit instability. More generally, any form of discrimination can adversely affect women's human capital investments and labour force attachment by lowering the market rewards to this behaviour (see also, Blau 1984; Blau and Ferber 1986; Ferber and Lowry 1976; and Weiss and Gronau 1981).

Conclusion

We have considered the contributions of neoclassical economic theory to our understanding of women's labour supply decisions, the gender division of labour within the family, and male–female differences in labour market outcomes. With the introduction of feedback effects, the separate strands of neoclassical theory analysing women's economic roles in the family and their labour market outcomes may be more tightly woven together. The causation runs not only from women's roles within the family to their resulting economic success, as human capital theorists emphasize, but also from their treatment in the labour market to their incentives to invest in market-oriented human capital and to participate in the labour force continuously. Thus, even a small amount of discrimination at an early stage of the career can have greatly magnified effects over the work life. While it is unlikely that labour market discrimination created the traditional division of labour between men and women in the family, it could certainly help to perpetuate it.

However, it is also the case that increasing opportunities for women in the labour market create powerful incentives to reduce gender differences in family roles and labour market behaviour. At the same time, women's increased attachment to the labour force, due not only to

these increased opportunities but also to changes in household technology and in tastes, may be expected to increase their market productivity and hence their earnings directly, and also to reduce statistical discrimination against them. Similarly, the movement of women into traditionally male jobs has the potential not only to increase the wages of those who become so employed, but to reduce overcrowding and increase wages in female jobs as well. Thus, just as a fuller understanding of the interrelationships between women's roles in the family and their status in the labour market helps us to understand the persistence of gender inequality in economic outcomes, it also enables us to appreciate how changes in either one of these spheres, or both, can induce a mutually reinforcing process of cumulative change. Recent signs of progress in reducing the pay gap in many of the advanced industrialized countries may well signal the beginnings of such a process.

In our emphasis upon the interdependence of women's status within the family and the labour market, we have in some respects returned to our starting point, for this conclusion bears a close resemblance to the views of the 19th-century observers which we reviewed at the outset. However, it is also clear that neoclassical economic theory has enhanced our understanding of the causes of gender differences in both the family and the labour market, as well as allowing us to comprehend better the links between the two sectors.

See Also

- Discrimination
- Family
- Household Production
- Housework
- Inequality Between the Sexes
- Labour Market Discrimination
- Labour Supply of Women
- Occupational Segregation
- Women and Work
- Women's Wages

Bibliography

Aigner, D., and G. Cain. 1977. Statistical theories of discrimination in labor markets. *Industrial and Labor Relations Review* 30(2): 175–187.

Arrow, K. 1973. The theory of discrimination. In *Discrimination in labor markets*, ed. O. Ashenfelter and A. Rees. Princeton: Princeton University Press.

Arrow, K. 1976. Economic dimensions of occupational segregation: Comment I. *Signs* 1(3): 233–237, Part II.

Becker, G. 1957. *The economics of discrimination*, 2nd ed. Chicago: University of Chicago Press, 1971.

Becker, G. 1965. A theory of the allocation of time. *Economic Journal* 75: 493–517.

Becker, G. 1981. *A treatise on the family*. Cambridge, MA: Harvard University Press.

Becker, G. 1985. Human capital, effort, and the sexual division of labor. *Journal of Labor Economics* 3(1): 533–558.

Bergmann, B. 1974. Occupational segregation, wages and profits when employers discriminate by race or sex. *Eastern Economic Journal* 1: 103–110.

Bergmann, B., and W. Darity Jr. 1981. Social relations in the workplace and employer discrimination. In *Proceedings of the thirty-third annual meeting of the Industrial Relations Research Association*, ed. B.D. Dennis, 155–162. New York: Industrial Relations Research Association.

Blau, F. 1984. Discrimination against women: Theory and evidence. In *Labor economics: Modern views*, ed. W. Darity. Boston: Kluwer-Nijhoff Publishing.

Blau, F., and M. Ferber. 1986. *The economics of women, men, and work*. Englewood Cliffs: Prentice-Hall.

Brown, C. 1984. *Consumption norms, work roles, and economic growth*. Paper presented at the conference on Gender in the Workplace. Washington, DC: Brookings Institution, November.

Doeringer, P., and M. Piore. 1971. *Internal labor markets and manpower analysis*. Lexington: D.C. Heath and Co.

Edgeworth, F. 1922. Equal pay to men and women for equal work. *Economic Journal* 32: 431–457.

Engels, F. 1884. *The origin of the family, private property and the state*. New York: International Publishers, 1972.

Fawcett, M.G. 1918. Equal pay for equal work. *Economic Journal* 28: 1–6.

Ferber, M., and B. Birnbaum. 1977. The 'new home economics': Retrospects and prospects. *Journal of Consumer Research* 4(1): 19–28.

Ferber, M., and H. Lowry. 1976. The sex differential in earnings: A reappraisal. *Industrial and Labor Relations Review* 29(3): 377–387.

Gilman, C. 1898. *Women and economics: A study of the economic relation between men and women as a factor of social evolution*. New York: Harper & Row, 1966.

Hartmann, H. 1976. Capitalism, partriarchy and job segregation by sex. *Signs* 1(3): 137–169, Part II.

Madden, J. 1973. *The economics of sex discrimination*. Lexington: D.C. Heath and Co.
Marx, K. 1867. *Capital: A critique of political economy*, vol. I. New York: International Publishers, 1967.
Mill, J.S. 1869. *The subjection of women*, 4th ed. London: Longmans, Green, Reader & Dyer, 1878.
Mincer, J. 1962. Labor force participation of married women. In *Aspects of labor economics*, National Bureau of Economic Research. Princeton: Princeton University Press.
Mincer, J., and S. Polachek. 1974. Family investments in human capital: Earnings of women. *Journal of Political Economy* 82(2): S76–S108, Part II.
Phelps, E. 1972. The statistical theory of racism and sexism. *American Economic Review* 62(4): 659–661.
Piore, M. 1971. The dual labor market: Theory and implications. In *Problems in political economy: An urban perspective*, ed. D. Gordon. Lexington: D.C. Heath and Co.
Polachek, S. 1981. Occupational self-selection: A human capital approach to sex differences in occupational structure. *Review of Economics and Statistics* 63(1): 60–69.
Treiman, D., and H. Hartmann (eds.). 1981. *Women, work, and wages: Equal pay for jobs of equal value*. Washington, DC: National Academy Press.
Webb, B. 1919. *The wages of men and women: Should they be equal?* London: Fabian Bookshop.
Weiss, Y., and R. Gronau. 1981. Expected interruptions in labour force participation and sex-related differences in earnings growth. *Review of Economic Studies* 48(4): 607–619.

Gender and Academics

Donna K. Ginther, Shulamit Kahn and Jessica McCloskey

Abstract

Although women have reached parity and surpassed men in the attainment of bachelor's degrees (Goldin et al. J Econ Perspect 20(4): 133–156, 2006; Ceci et al. Psychol Sci Public Interest 15(3): 75–141, 2014), their representation within academic departments and disciplines depends on the field and rank. Here, we review the literature about women in academia, focusing on the evidence from the economics literature, but supplementing it with notable studies from other disciplines. We also examine the special case of the economics profession, where – surprisingly – women's progress has stagnated.

We start by describing the representation of women in science academia and its antecedents in higher education. Since, in mathematics-intensive sciences, the under-representation has its roots prior to the doctorate, we briefly summarise what is known about gender differences related to mathematics and science at earlier ages. In particular, we examine the impact of role models, bias and stereotype threat in explaining the differences. We then transition to research on gender differences in academic career outcomes, considering issues related to work–life balance and bias in the academic hiring process, in academic productivity, in promotion and in salaries. Finally, we discuss how policies influence the representation of women in academia.

Keywords

Academia; Division of labour; Equality; Gender; Parity; Pay gap; Wage gap; Work

JEL Classifications

J16; J31; J7

Although women have reached parity and surpassed men in the attainment of bachelor's degrees (Goldin et al. 2006; Ceci et al. 2014), their representation within academic departments and disciplines depends on the field and rank. Here, we review the literature about women in academia, focusing on the evidence from the economics literature, but supplementing it with notable studies from other disciplines. We also examine the special case of the economics profession, where – surprisingly – women's progress has stagnated.

We start by describing the representation of women in science academia and its antecedents in higher education. Since, in mathematics-intensive sciences, the under-representation has its roots prior to the doctorate, we briefly summarise what is known about gender differences related to mathematics and science at earlier

ages. In particular, we examine the impact of role models, bias and stereotype threat in explaining the differences. We then transition to research on gender differences in academic career outcomes, considering issues related to work–life balance and bias in the academic hiring process, in academic productivity, in promotion and in salaries. Finally, we discuss how policies influence the representation of women in academia.

What are the Numbers?

Figure 1a shows the percentage of women among higher education faculty in the USA from the National Center for Education Statistics (IPEDS) for the years 2006 through 2011, for all fields combined. Women hold only about one-third of tenured faculty positions and somewhat less than half of tenure-track untenured faculty positions. On the other hand, women are substantially over-represented among non-tenured, non-tenure-track faculty (since all faculty together average more than 50%).

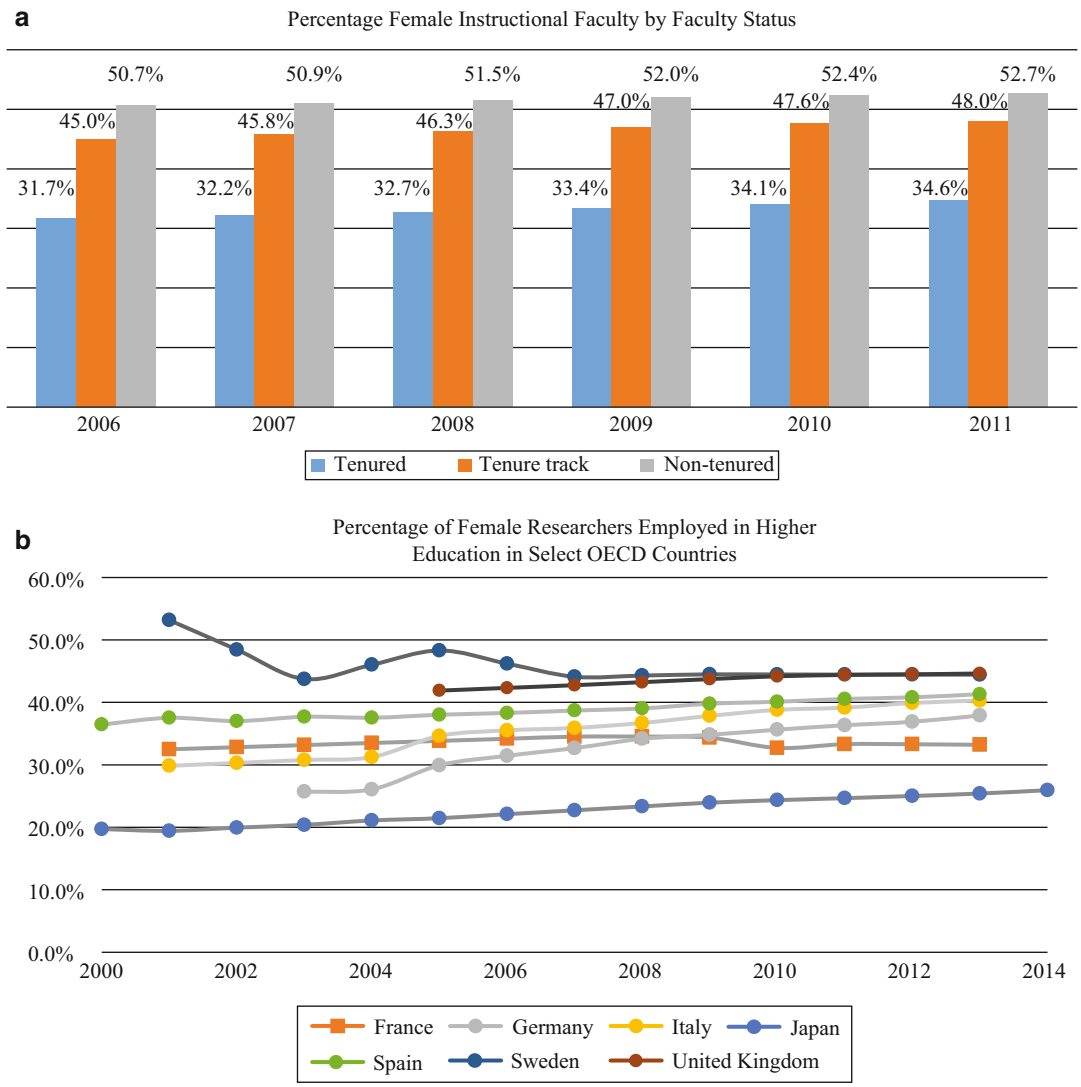
Figure 1b shows the percentage of women among researchers in higher education in selected OECD countries from 2000 to 2014 for all fields combined. As of 2014, women made up 45% of all researchers in higher education in the UK and Sweden, 41% in Spain, 40% in Italy, 38% in Germany, 33% in France and 25% in Japan. In both the USA and these countries, the representation of women in academia has been remarkably stable over the past decade.

It would be ideal to have longer time-trends and field-specific data. Unfortunately, we can only do this comprehensively for science, technology, engineering and mathematics (STEM) disciplines, where considerable data is available for the USA from the National Science Foundation. As shown in Fig. 2a, while the percentage of women has been increasing in all fields, women exceed 50% of tenure-stream academics only in psychology. They are at 43% in social sciences (excluding economics) and 30% or less in all other STEM fields. For humanities, we can only use labour force surveys that do not distinguish among faculty ranks. These indicate a constant 50% female among postsecondary humanities faculty in the 1990s and 2000s (see http://www.humanitiesindicators.org).

Figure 2b shows the percentage of women researchers in the natural sciences in selected OECD countries from 2000 to 2014. With the exception of Sweden, we observe an increase in the representation of women between 2000 and now. The 2013/2014 percentages of women in Sweden, the UK, Spain and Italy are remarkably similar at 38–41%. In Germany and Japan, the percentages of women have increased over the past decade and a half, but are currently only 33% and 22% respectively.

Were the fields that are underrepresented in US academia similarly under-represented among US PhDs? Figure 3a shows that the percentage of women among PhDs has increased dramatically in all fields over this period. Women currently hold the majority of PhDs granted in Life Sciences, Psychology, Other Social Science (except economics) and Humanities, and more than 30% in other STEM fields, except for engineering and computer science. These numbers are clearly higher than the current female percentage among tenured faculty. However, that is an incorrect comparison. Tenured faculty 2006–2011 would have received their PhDs in the 1970s, 1980s and 1990s, when the average percentage of women among PhDs was 27%. Comparing this percentage to the approximately one-third women among tenured faculty suggests that women PhDs were equally or somewhat more likely than men to become tenured academics. Similarly, tenure-track untenured faculty 2006–2011 would have received PhDs between 1995 and 2010, when the percentage of women among PhDs awarded averaged 39%. This suggests that women PhDs during this period were actually *more* likely than men to be in tenure-track jobs.

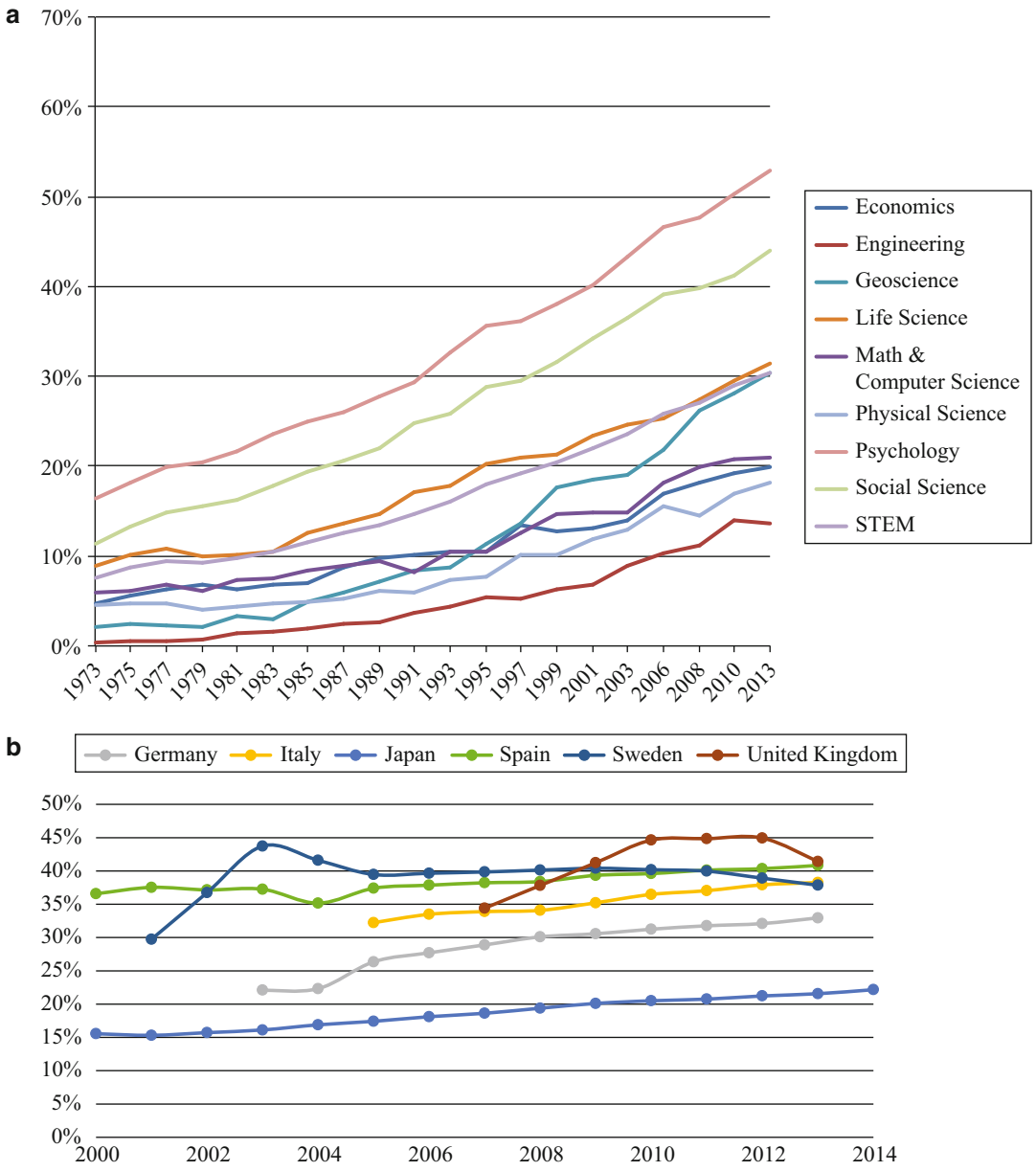
For the sciences, we can more accurately measure whether women and men PhDs proceed to tenure-stream jobs at similar rates. Using the data in Fig. 2, in Ceci et al. (2014) we matched field-specific rates of PhDs to their rate of holding tenure-stream jobs seven years later, and found that in mathematics-intensive sciences the proportion of PhDs who entered tenure-track jobs was



Gender and Academics, Fig. 1 (**a**) Percentage of women in academic positions in the USA 2006–2011 (Source: Integrated Post-Secondary Data System); (**b**) percentage female researchers in higher education in selected OECD Countries 2000–2014 (Source: UNESCO 2000–2014)

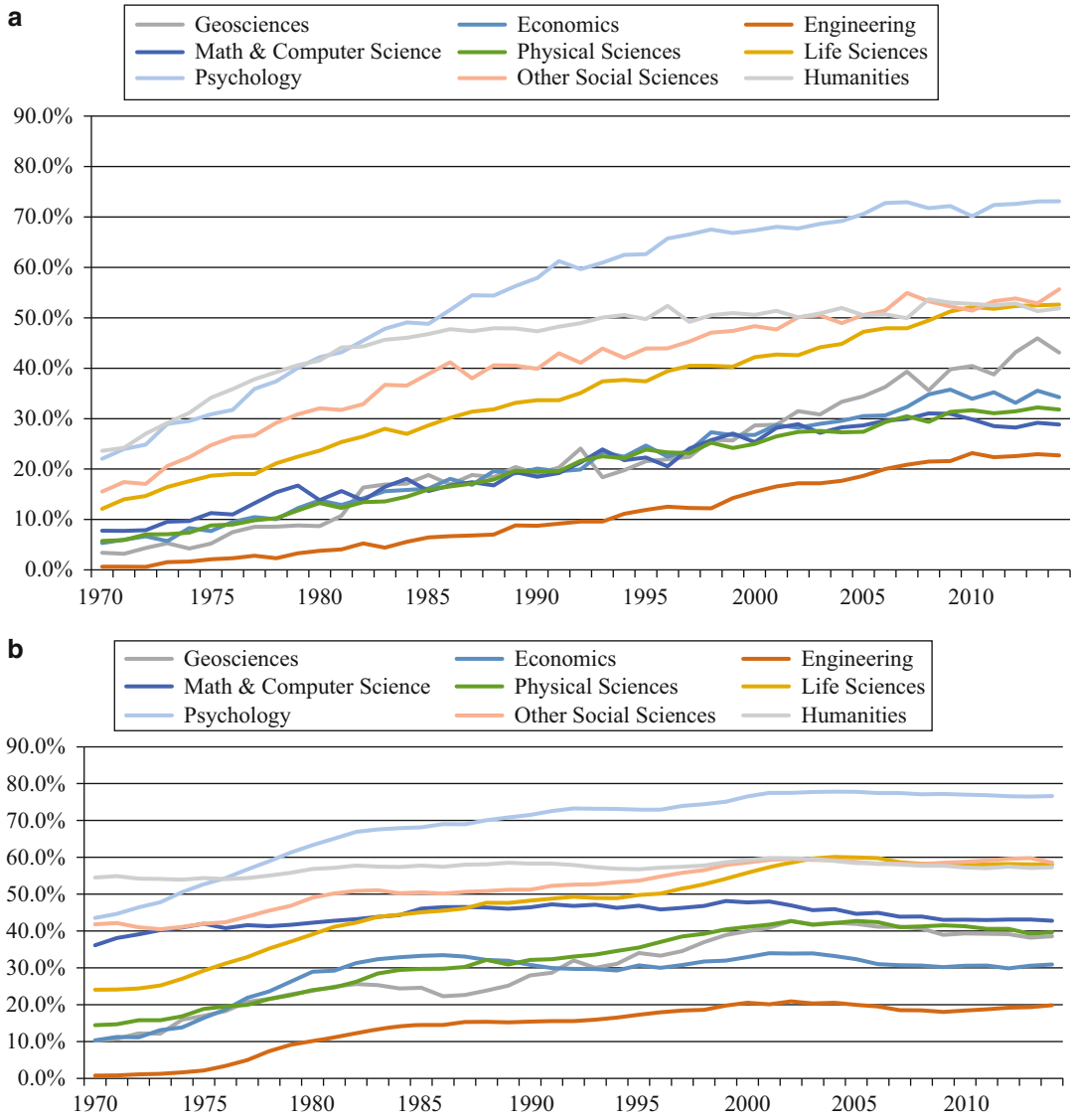
similar for men and women; however, for the life and behavioural sciences (life, psychology, social sciences excluding economics), lower percentages of women than men entered academic tenure-track jobs. For humanities as a whole, the constant 50% of women among postsecondary faculty in the 1990s and 2000s was quite similar to the percentage of women among PhDs granted in the 1990s and 2000s and higher than the percentage of women among PhDs granted in earlier decades. We thus conclude that progression from PhD to tenure-track and tenured jobs is currently similar for men and women for all fields except the life and behavioural sciences.

These findings suggest that underrepresentation in more mathematics-intensive fields starts earlier in people's lives. Figure 3b shows the gender breakdown by discipline in bachelor's degrees in the USA. Starting in the 1980s and 1990s, women received the majority



Gender and Academics, Fig. 2 (**a**) Percentage of women among tenure-stream faculty in the natural and behavioural sciences (Source: Survey of Doctorate Recipients, 1973–2013); (**b**) percentage of women among higher education researchers in the natural and behavioural sciences for selected OECD Countries (Source: UNESCO, 2000–2014)

of bachelor's degrees in psychology, humanities, social sciences (excluding economics) and life sciences. They also greatly increased their share of bachelor's degrees in other sciences, and since 2000 have earned over 40% of bachelor's degrees in all disciplines except engineering, economics and computer science. Goldin et al. (2006) found that much of this relative increase in women's college completion rate was attributable to the improvement in college preparation of girls



Gender and Academics, Fig. 3 (**a**) Percentage of PhDs conferred on women in the USA, 1970–2014 (Source: IPEDS); (**b**) percentage of bachelor's degrees conferred on women in the USA, 1970–2014 (Source: IPEDS)

relative to boys (especially in STEM), which itself was probably due to women's increased expected returns from going to college.

Are men and women equally likely to proceed from bachelor's degrees to PhDs, assuming a seven-year time gap? The percentage of women among PhDs was lower than for corresponding bachelor's degrees in the earlier decades shown, but recently this gap has narrowed in all fields (also found by Chiswick et al. 2010). Most recently, the percentage of women among PhDs versus bachelor's degrees seven years earlier was the same for mathematics-intensive sciences, but lower in the humanities and life and behavioural sciences (in humanities, compare 52% women among PhDs to 63% women among bachelor's degrees seven years earlier; in the life and behavioural sciences, compare 58% among PhDs but 65% among bachelor's degrees).

Where did these women go instead? In the life and behavioural sciences, Ceci et al. (2014) found that more women than men had master's degrees (27% v. 21.8%) and professional degrees (9.1% v. 7.9%). However, fewer women than men (38.2% v. 45.5%) stopped their education after attaining a bachelor's degree.

The numbers lead us to conclude that in the USA there is a larger drop-off of women than men in the transition from BA to PhD in the fields where women are more common (humanities, life and behavioural sciences), but not in mathematics-intensive fields. There is no gender difference in the transition from PhD to tenure-track academia in mathematics-intensive sciences and humanities, whereas in life and behavioural sciences the drop-off from PhDs to tenure-stream academia is greater for women.

Since the under-representation of women in mathematics-intensive STEM fields has its roots even earlier than college, we next describe the literature on some factors explaining this underrepresentation.

Gender Differences in K-12 and Undergraduate Educational Outcomes

Mathematics is considered the gatekeeper to careers in STEM disciplines (Ceci et al. 2014; Lavy and Sand 2015). On average, in the USA girls score better than boys in mathematics in some grades (4–9) but not in high school, although high school average gaps have dropped rapidly to less than one-tenth of a standard deviation (Hyde and Mertz 2009); gender differences at the top tail of the high school mathematics distribution have also dropped rapidly (from 1:13.5 to 1:3.8; Wai et al. 2010), but remain high. Yet research shows that environmental factors and context play a role in gender differences in mathematics performance. Ellison and Swanson (2010) found variation in the gender gap across schools. Pope and Sydnor (2010) found state geographic variation in gender test gaps, at the state level. Else-Quest et al. (2010), Penner (2008) and others found trans-national variation.

One explanation for the differences in girls' and boys' early mathematics and science attainment relates to instructor gender. For middle school students, Dee (2005, 2007) found that assignment to a same gender teacher improved both boys' and girls' achievement as well as teachers' perception of students and students' engagement, *in all subjects*. Ehrenberg et al. (1995) found that same-gender teachers did not affect learning but influenced teachers' subjective evaluation of students in mathematics, science and reading. Antecol et al. (2012) found marginal positive effects on female students' mathematics scores only for female instructors with strong mathematical backgrounds, no effect on their reading scores, and no impact of male instructors on male students at all. More recently, Lavy and Sand (2015) found that girls in Israel with elementary and middle school teachers biased against girls in mathematics took fewer high school mathematics and science courses and were less likely to major in mathematics and science in college or work in STEM.

Same-gender role model effects extend to college. Using randomly assigned students, Carrell et al. (2010) found that female instructors in male-dominated STEM fields improved female students' performance in mathematics and science classes and the likelihood of taking future STEM classes and majoring in STEM, with the results greatest for top students. Based on a natural experiment, Griffith (2014) found that same-gender instructors improved students' performance only in fields traditionally dominated by the opposite gender, but had no effect on major choice or course-taking behaviour. Observational studies – without random assignment of students – have mostly found that female instructors improved female students' outcomes (Rask and Bailey 2002; Hoffman and Oreopoulos 2009; Bettinger and Long 2005), but did not affect the registration choices of most students (McGoldrick and Schuhmann 2002; Canes and Rosen 1995). Ashworth and Evans (2001) found that female students were more likely to study economics when there was a critical mass of other female students and/or a female teacher.

Female role models were also important in observational studies at the graduate level. In economics, Hale and Regev (2014) found a positive correlation between the number of female faculty and the number of female graduates six years later, suggesting that women graduate students were attracted to and/or encouraged by women faculty. Dolado et al. (2012) found greater shares of women in a given economics sub-field to be correlated with greater probability of women later choosing that field.

Research has also examined how gender differences in response to competition play a role in mathematics-related outcomes. Niederle and Vesterlund (2010) argue that gender differences in mathematics test scores may indicate different responses to competitive pressures associated with test-taking. Cotton et al. (2013) find results consistent with that argument in five sequential mathematics contests among elementary-school children. Boys scored higher in the first round than girls, but only when there was time pressure. Girls scored better in later rounds. Landaud et al. (2016) found that girls enrolled in to more competitive high schools in France were significantly less likely to choose a high school mathematics or science major.

In economics, some have argued that teaching methods decrease female interest in the subject. Bansak and Starr (2010) found that students viewed economics as a business-oriented field that emphasised mathematical skills and money-making, which decreased women's interest relative to men's. Similarly, Lewis and McGoldrick (2001) argue that reformulating standards might allow for a more inclusive classroom. A current randomised trial headed by Claudia Goldin is experimenting with multiple interventions associated with mentoring of female students and curriculum changes in order to increase the number of women majoring in economics (http://scholar.harvard.edu/goldin/UWE).

A final gender difference in the decision to get a PhD relates to the macroeconomy. Bedard and Herman (2008) found that women's decisions to attend graduate school were acyclical, while men's decisions were counter-cyclical, so that when macroeconomic conditions worsened, the lower opportunity cost of attending graduate school increased men's (but not women's) enrolment. Chiswick et al. (2010) also found that men's doctorate enrolment increased with unemployment. Conley et al. (2016) found that men who entered economics graduate school in periods with few outside opportunities (high unemployment) later had higher research productivity, but women who entered then had lower research productivity, and offered a similar cyclical selection explanation.

Before leaving the education topic, we note that Leslie et al. (2015) tried to link PhD attainment to faculty attitudes in the discipline, finding that in disciplines where expectations of brilliance are viewed as the key to success – as opposed to hard work – women were less likely to obtain doctorates. However, Ginther and Kahn (2015) show that once the mathematics requirements of a particular discipline are included in their analysis, these expectations have no explanatory power.

Thus, the association between gender norms, role models and mathematics/STEM plays a role in determining educational outcomes and choices from middle school to PhD, giving rise to the observed gender differences in academic careers. We next turn to how women fare once they enter tenure-stream academic jobs.

Gender Differences in Tenure-Stream Positions

Hiring

As described above, the proportion of women with tenure-track positions in the life and behavioural sciences is lower than might be expected based on the number of doctorate degrees awarded. Was this due to women not being offered academic positions, or to their choices to opt out of academic positions?

Using data from the department chairs of six STEM departments at Research I universities from the early 2000s, a National Research Council study (2010) indicated that, conditional on applying, women were more likely to get an interview and more likely to receive job offers in all six departments. Also, the percentage of applications

from women was consistently lower than the percentage of PhDs earned by women. These same results were found for both assistant tenure-track positions and more senior tenured positions.

This does not necessarily rule out bias in the interview and hiring process, since if on average women applicants are more qualified than male applicants, the proportion of women receiving interview and job offers might understate bias. Recent experimental studies on the role of bias in potential hires have produced contradictory results. In a relatively small sample, Moss-Racusin et al. (2012) found that science faculty evaluating hypothetical identically qualified graduate students evaluated the men as more competent; they were more likely to be hired as well as being given higher starting salaries. Williams and Ceci (2015) found the opposite, also in an experimental setting but with a larger sample. They had faculty evaluate hypothetical equally qualified male and female applicants for assistant professor positions in biology, engineering, economics and psychology at different institution types nationwide. In most cases, both male and female faculty preferred female applicants over identically qualified males with matching lifestyles. The exception, showing a male preference, was male economists. The average preference for women was significant within five of six categories of family status (e.g. married without children).

Outside the USA, Krause et al. (2012) conducted an experiment randomly assigning applications of PhD economists for a postdoctoral position at a European research institute to a treatment group whose applications removed information on name, age, gender and nationality versus controls with this information included. In the control group, but not in the treatment group, women applicants received more interviews than men. Using the difference in French teacher accreditation exam scores between written (gender-unknown) and oral (gender-known) as a natural experiment, Breda and Hillion (2016) showed that the gender under-represented in that field was systematically favoured when gender was known.

Thus, all in all, there is little evidence of bias against women and some indication of bias towards women in the hiring process in academia when the person's record is known.

Opting Out

Clearly, women's decision not to pursue tenure-stream positions affects their representation in academia. Evidence of fewer applications yet slight advantages in hiring and interviewing are consistent with the argument that relatively more women are opting out of academia. Ley and Hamilton (2008) examined women's attrition in biomedical sciences at US medical schools. A roughly equal share of women were admitted to medical schools (51%) and working as instructors at medical schools (49%). The percentage of women dropped, however, at later stages of the traditional academic career track (39% assistant level, 25% associate level, 17% full level). They found that women in biomedical fields were not applying for NIH funding at the independent research stage (in between the postdoc and a tenure-stream appointment). Ginther and Kahn (2009) looked at the probability of STEM PhDs holding a tenure-track job within nine years of graduating. They found that married women with children were significantly less likely to take tenure track positions. Ginther and Kahn (2015) repeated this analysis for social and behavioural science fields, finding very similar results. Wolfinger et al. (2008) also found that women with children during the first five years post-PhD are considerably less likely than men to choose tenure-track jobs.

Institutional factors may also play a role in the gender diversity of the faculty. Ehrenberg et al. (2012) found that having more women in high-ranking administrative positions (trustees, presidents/chancellors and provosts/academic vice-presidents) was associated with having more women on the faculty between 1984 and 2007, with the largest gains appearing at smaller institutions.

Economics

Economists have examined gender differences in jobs after the PhD. Chen et al. (2012) report that compared to males, female candidates were more likely to be in government or private sector jobs

and less likely to end up in academic jobs. Hilmer and Hilmer (2007) found that females with male advisors were more likely to accept research-oriented first jobs than males with male advisors. They found no significant difference between females working with male versus female advisors.

Productivity

Once in academic positions, productivity – measured by publications, citations and research funding – is key to securing tenure and remaining employed in the academy. Across academic fields, almost all research shows that women write fewer papers, but on average have the same number of citations per paper (see Ceci et al. (2014) for a review of the literature). There is some evidence that gender differences in productivity are converging (Borrego et al. 2010). Gender differences in productivity have often been cited as the leading explanation for gender differences in salaries in the general labour market (Altonji and Blank 1999). Economists have probed whether these productivity differences are due to gender differences in time use, the impact of having children, professional networks, number of co-authors, access to institutional resources and support, and likelihood of specialising.

Women may be less productive because they devote less time to work (Bellas and Toukoushian, 1999). Ceci et al. (2014) found that women and men in STEM tenure-stream positions work the same number of hours. However, women with children published significantly fewer papers than men with children in geoscience, economics, physical, and life and social science disciplines. In contrast, there were no significant gender gaps in publications for women and men without children in life science or social science – suggesting that time devoted to caring for one's family may contribute to the gender gap in publications. Krapf et al. (2014) compared the research productivity of economists and found a negative effect of parenthood for unmarried mothers, and a positive impact for unmarried fathers. They also found evidence that becoming a mother before the age of 30 had a negative impact on women's research productivity. Joecks et al. (2014) examined 400 researchers in business and economics in Austria, Germany and Switzerland. They found evidence that only the most productive mothers self-select into academic research careers.

Time use was also a factor in Manchester and Barbezat's (2013) study of economics faculty. There, gender differences in both time allocation (division of time between research and other duties) and time concentration (distribution of time during the academic year relative to summer) contributed to women submitting fewer papers, with concentration being most important.

Non-research obligations may also influence research productivity. Taylor et al. (2006) found that teaching and service have significant negative impacts on research productivity of academic economists. Harter et al. (2010) found that in the USA, male economics faculty – particularly at the assistant professor level in research universities – spent less time on teaching and more time on research than female faculty.

Women may also be less productive because of fewer resources. Duch et al. (2012) showed that fields that required significant research resources (such as molecular biology) also had a larger gender gap in publications. However, gender differences in research awards are negligible. Ginther et al. (2016) and Ley and Hamilton (2008) find that women are equally or somewhat more likely to receive NIH R01 Type 1 research awards; however, women are disadvantaged in receiving additional funding of the same research topic – NIH R01 Type 2 research awards (Ley and Hamilton 2008). Furthermore, women submit fewer research proposals than men (Ginther et al. 2016; Ley and Hamilton 2008). Sege et al. (2015) found women researchers in a major medical school had less start-up support than men.

Gender differences in co-authorship contributes to gender differences in productivity in economics. Hamermesh (2013) has noted the increasing importance and reliance on co-authorship in economics profession. Others have found that co-authorship among economists appeared to increase the overall production of articles for both men and women (Maske et al. 2003; Cainelli et al. 2015, 2012). Research

shows that economists tend to co-author with those in their gender (McDowell and Smith 1992; McDowell et al. 2006; Boschini and Sjogren 2007). Given the under-representation of women in the economics profession, this would provide one potential explanation for why women publish fewer papers, at least in economics.

However, email and internet technology may level the playing field. Butler and Butler (2011) found that for academics in political science, technological change led women to increase their rate of co-authorship faster than men in the 1990s and made women more willing to take jobs at smaller departments because collaboration across universities was more possible. Similarly, Ding et al. (2010) found that IT availability increased research output and co-authorships for women at non-elite institutions, more than for men or for both genders at elite institutions.

Biased evaluations of work could also play a role in differences in publication numbers. However, research shows no gender differences in journal acceptance rates in economics (Blank 1996; Abrevaya and Hammermesh 2012) nor in other disciplines (Ceci et al. 2014). An experiment found no effect of blind review on gender differences in acceptance rates for a Swedish economics conference (Carlsson et al. 2012).

Thus we find that women publish fewer papers than men, and these productivity differences are associated with the presence of children, time use during and across the academic year, research funding and co-authorship patterns. Technology has mitigated some of the co-authorship disadvantage, but women still lag behind men in this important measure of academic careers.

Promotion

In the USA, gender differences in academic promotion depend upon the field of study. Ginther and Kahn (2009) found that after controlling for research productivity and other factors, women were equally likely to receive tenure in physical science and engineering fields, but not life sciences. Ginther and Kahn (2015) found that women were significantly less likely to receive tenure in economics, but not other social sciences, and significantly less likely to be promoted to full professor in economics, sociology and linguistics. In earlier work, women were less likely to be promoted in the humanities (Ginther and Hayes 1999, 2003). McDowell et al. (1999, 2001) found promotion prospects significantly improved for female economists by the end of the 1980s. However, Kahn's (1993, 1995) results found the opposite, and examining more recent data (Ginther and Kahn 2004, 2015) found large promotion gaps for women in economics.

Academic promotion differs considerably across countries. In several cases, promotion differences were due to productivity. Schulze et al. (2008) found that gender and children did not matter for the probability of being tenured after controlling for productivity in Germany, Austria and the German-speaking part of Switzerland. Groeneveld et al. (2012) found that in a large Dutch university, academic women's lower promotion rates were explained by years of service and external mobility. Lissoni et al. (2011) found that Italian academic women are as likely to be promoted as men with similar publication records. Danell and Hjerm (2013) found that women were significantly less likely than men to become full professors in Sweden, but less so among those who had previously held postdoctoral fellowships, suggesting that promotion may reflect ability.

In other countries, promotion differences remain even after controlling for productivity. Takahashi and Takahashi (2015) found that in Japan, women were substantially more likely to remain in lower-level lecturer positions. At higher levels they found women with children were less likely than comparable men to be promoted from associate professor to full professor, but single childless women were more likely to be promoted. Examining women in the UK, Ward (2001b) found that even after controlling for career breaks and publication history, male academics are more likely to be promoted.

Results for France are mixed. Lissoni et al. (2011) found that equally productive French women were less likely than men to be promoted. Similarly, controlling for productivity, Sabatier (2010) found that female biologists in France

were promoted significantly more slowly than males and that different factors affected promotion likelihood for men and women. Also in France, Bosquet et al. (2014) found that in a national competition for promotion of economists, gender has no significant effect on promotion, but women were significantly less likely to be candidates for promotion.

Austen (2004), Cooray et al. (2014) and Kahn (2012) found that similar Australian women academics were less likely to be promoted than men, although Kahn (2012) found that women were more likely to be promoted after taking workshops on applying for promotions. In Australia, faculty must apply for promotion, and Kahn (2012) argued that the earlier promotion gap was due to women's lower application rates.

Finally, applying for promotion was also key in Italy. De Paola et al. (2015) examined the multistep Italian promotion system and found that women and men were equally likely to score well on the (anonymous) qualifying exam, but that qualified women were significantly less likely to apply for open positions than men. In Italy and Spain, there is also a (non-anonymised) oral exam by a randomly assigned evaluation committee. De Paola et al. (2015, 2016), Bagues et al. (2015) and Zinovyeva and Bagues (2011) find conflicting results on whether the evaluation committee gender composition leads to most favourable results for women.

In sum, we find mixed results on promotion. In some fields – primarily economics and life sciences, and in some countries including Japan, the UK and perhaps France or Australia – women are less likely to be promoted than men. Some, but not all, of this gap can be explained by gender differences in productivity or in applying for promotion. In the USA, economics is one field where women are significantly less likely to be promoted than men at all levels, even after controlling for the publication record.

Salary

Women are paid less than men in academia (https://www.aaup.org/our-work/research/annual-report-economic-status-profession; Toutkoushian et al. 2007), and this has been documented extensively over time (Barbezat 1987a, b; Broder 1993; Ferber and Kordick 1978; Gordon et al. 1974; Robinson and Monks 1999). Factors used to explain the gender gap in salaries include field and academic rank, productivity, parenthood and returns to seniority/monopsony.

Field and Rank

Field and rank are the most important explanations of salary gaps, because men are concentrated in higher-paying fields and are concentrated in the higher ranks (Ginther 2004). Failure to control for those factors will result in an overstated salary gap. Similarly, teaching-intensive institutions pay less than research-intensive institutions. Ginther and Hayes (1999, 2003) found no gender difference in salaries in the humanities within academic ranks. International evidence also points to the importance of field and rank. Warman et al. (2010) found that the gender earnings gap at Canadian universities had narrowed, and the bulk of the remaining gender gap could be explained by differences in men's and women's rank and field. Kaszubowski and Wolszczak-Derlacz (2014) found that gender differences in salary were mostly due to academic rank in Polish academia.

Productivity

Rank is endogenous and could be due to lower academic productivity of women. Hilmer et al. (2012) found that in doctoral-granting economics departments in large public universities in the USA, research influence (measured by citations) was a strong predictor of salary, as was departmental prestige. After controlling for these factors, they found no significant impact of gender on salaries. Ward (2001a) and Euwals and Ward (2005) found that in the UK, time out of the profession results in a large financial penalty, and that career gaps along with productivity could explain the gender salary gap.

Parenthood Penalty

As discussed above, having children might decrease salaries due to lower productivity, or for other reasons. Manchester et al. (2010, 2013) examined the impact of stopping the tenure clock

on both promotion and salaries. They found that stopping the tenure clock had no impact on promotion, but did result in a significant salary penalty. In recent work, Kahn and Ginther (2016) found that marriage and children have less of a negative impact on women's STEM academic salaries than for women with the same degrees working outside academia.

Monopsony

In many cases a given geographic location has only one university, and that university holds monopsony power over its current faculty. Several researchers have found that the returns to seniority are negative for faculty (salary inversion) (Ransom 1993; Hallock 1995; Bratsberg et al. 2010; Brown and Woodbury 1998); however, Barbezat and Donihue (1998) found the opposite. Monopsony power can exacerbate gender salary differences if men are more likely to receive outside offers than women or women are less likely to move. Hilmer and Hilmer (2010) found that the seniority penalty for women economists was nearly double that of men, and that men earn higher salaries with each move while women's salaries only increase with two or more moves. Barbezat and Hughes (2005) found that women experienced an 8% salary penalty for moving to a second job. In the UK, Blackaby et al. (2005) found a within-rank gender pay gap among academic economists that they suggest may be due to women's lower likelihood of receiving outside offers.

Remaining Gaps

In other cases, while controls for the above factors narrowed the salary gap, they did not erase them. In the natural sciences, Ginther (2004) found small salary gaps at the assistant professor rank that grew for the associate and full professor ranks. These gaps were not fully explained by field, marital status, children or productivity. Takahashi and Takahashi (2015) found that Japanese women economists were paid significantly less within academic ranks, despite rigid pay schedules. Some of the pay gap may result from women being hired at lower wages when they start (Toumanoff 2005). Of particular concern are gender differences in evaluation: in a study of a large US public research university, Carlin et al. (2013) found that both subjective and objective productivity measures increased men's salaries, but did not increase women's. Finally, sometimes, gender pay gaps were more complicated. At a large US public university, Binder et al. (2010) found that, controlling for productivity, the largest gender salary gaps were in departments with low concentrations of women, suggesting that decentralised salary setting in departments may serve to depress women's salaries.

In sum, women choose lower-paid academic fields and are also more prevalent in the lower academic ranks, two factors that explain much of the overall gender salary gap in academia. Other choices by women, such as productivity and parenthood, serve to exacerbate the gender salary gap. That said, institutional factors exacerbate gender differences in salaries. Wage-setting institutions at the department level and the monopsonistic market faced by many in academia reinforce the gender wage gap.

Other Outcomes

We briefly mention a few additional outcomes. Mixon and Trevino (2005) found that women were significantly less likely to have a named professorship in economics departments in the US South. In Italy, Addis and Villa (2003) found that women were less likely to serve on the editorial boards of economics journals. In contrast, Donald and Hamermesh (2006) found that women were more likely to win American Economic Association elections.

Ceci et al. (2014) found that women's job satisfaction with academic careers converged with men's between 1997 and 2010, with the exception of social sciences and economics, where the gap grew and women were less satisfied. Bender and Heywood (2006) found that women's satisfaction in academic science matched men's. Ward and Sloane (2000) find that job satisfaction does not differ by gender in Scottish universities.

Potential Interventions

We have documented gender differences in productivity, promotion and salaries in academic careers. In some cases, these differences can be readily explained by the family, monopsony, resources, co-author networks and other factors; in other cases we cannot rule out gender differences in how women are treated and evaluated during their academic careers. Economists have begun addressing gender differences in academic careers through the CeMENT mentoring trial for junior economics faculty at research institutions. Starting in 2004, the CeMENT trial randomly assigned junior female economists to a mentoring treatment workshop or a control group without mentoring. An interim evaluation of CeMENT by Blau et al. (2010) showed that women in the treatment group published more papers, published more in the highest ranked journals and were more successful in obtaining federal research funding. Based on CeMENT's success, mentoring programmes have been started in the economics profession in Africa, China and Japan, as well as in academic philosophy (https://www.aeaweb.org/content/file?id=520).

Universities in the USA have adopted policies for parents to stop the tenure clock in the case of birth or adoption. Manchester et al. (2010, 2013) found that these policies had no impact on women's promotion, but had a negative impact on salaries at one Midwestern university. However, Antecol et al. (2016) found that, in the top 50 US economics departments, gender-neutral stop-clock policies reduced female tenure rates while significantly increasing male tenure rates.

Others have advocated for training in unconscious bias as a means of combatting gender differences in academic careers. Carnes et al. (2015) ran a randomised controlled trial of bias training at the University of Wisconsin-Madison and found that gender bias was reduced in treated departments. However, their study did not evaluate whether the reduction of gender bias influenced gender differences in academic outcomes such as hiring and promotion. In the private sector, this kind of bias training has not promoted diversity (Dobbin et al. 2012; Kalev et al. 2006), although Bohnet et al. (2015) found evidence that joint evaluation by a committee allowed evaluators to focus on performance and reduced gender bias.

Conclusions

The status of women in the academy depends critically on the country and academic field. Although in the USA women earn more than half of both bachelor's and PhD degrees awarded in the humanities, life, behavioural and social sciences (excluding economics), they are less likely than men to transition from bachelor's to doctorates to tenure-stream faculty positions in these fields. In contrast, we find that women have made significant gains in mathematics-intensive science doctorates and are somewhat more likely than men to transition from PhD to academic careers in those fields. The under-representation of women in mathematics-intensive fields has its roots prior to college, and much of it can be attributed to gender differences in role models and gender norms in mathematics.

The data and experimental evidence do not show evidence of bias against women in academic hiring. Instead, there is some evidence of a preference for more female faculty. Once in tenure-stream academic positions, however, women publish fewer papers, although they have the same number of citations per paper as men. Several inter-related factors contribute to women's lower productivity. Having children lowers the number of publications for both men and women, but more for women. Women devote less time to research than men, in part because they are more likely to be employed at teaching-intensive institutions. Co-authorship increases the rate of productivity, but women are more likely to co-author with other women and in fields like economics they will have fewer opportunities to do so. Resources may also matter – women are less likely to receive grant renewals in biomedical fields – and their productivity suffers in these relatively expensive disciplines.

Productivity is a key determinant of both promotion and salaries, but even after controlling for

productivity women are less likely in some fields (e.g. economics, life sciences) and in many countries (Japan, UK etc.) to be promoted. Women's pay also suffers relative to men's. Some of this can be explained by productivity, presence of children, field and academic rank. Yet there is also evidence that given similar levels of productivity, women's evaluations suffer, leading to lower salaries over time (Carlin et al. 2013). The fact that most academic employers are monopsonists can lead to significant gender salary differences if women are less likely to receive outside offers or less likely to move.

Stopping the tenure clock has not been shown to increase women's promotion rates, and instead may decrease it. When we combine this with the gender differences in how women's research is evaluated by their peers, then attention to evaluation and its implications for salary and promotion are warranted.

Interventions such as the CeMENT mentoring treatment have been successful at increasing women's productivity in the economics profession. It remains to be seen whether CeMENT will successfully narrow the gender promotion gaps in the economics profession. Although women have made considerable strides in academic careers, progress has been uneven across disciplines and countries.

See Also

- ▶ Gender Roles and Division of Labour
- ▶ Monopsonistic Discrimination and the Gender Wage Gap
- ▶ Women's Work and Wages

Bibliography

Abrevaya, J., and D. Hammermesh. 2012. Charity and favoritism in the field: Are female economists nicer (to each other)? *Review of Economics and Statistics* 94(1): 202–207.

Addis, E., and P. Villa. 2003. The editorial boards of Italian economics journals: Women, gender, and social networking. *Feminist Economics* 9(1): 75–91.

Altonji, J., and R. Blank. 1999. Race and gender in the labor market. In *Handbook of labor economics* 3(C), 3143–259.

Antecol, H., O. Eren, and S. Ozbeklik. 2012. The effect of teacher gender on student achievement in primary school: Evidence from a randomized experiment. IZA Discussion Paper No. 6453.

Antecol, H., K. Bedard, and J. Stearns. 2016. Equal but inequitable: Who benefits from gender-neutral tenure clock stopping policies? IZA Discussion Paper No. 9904.

Ashworth, J., and J. Evans. 2001. Modeling student subject choice at secondary and tertiary level: A cross-section study. *Journal of Economic Education* 32(4): 311–320.

Austen, S. 2004. Gender differences in academic rank in Australian universities. *Australian Bulletin of Labor* 30(2): 113.

Bagues, M., M. Sylos-Labini, and N. Zinovyeva. 2015. Does the gender composition of scientific committees matter? IZA Discussion Paper No. 9199.

Bansak, C., and M. Starr. 2010. Gender differences in predispositions towards economics. *Eastern Economic Journal* 36(1): 33.

Barbezat, D. 1987a. Salary differentials or sex discrimination? Evidence from the academic labor market. *Population Research and Policy Review* 6: 69–84.

Barbezat, D. 1987b. Salary differentials by sex in the academic labor market. *Journal of Human Resources* 22(3): 443–455.

Barbezat, D., and M. Donihue. 1998. Do faculty salaries rise with job seniority? *Economics Letters* 58(2): 239–244.

Barbezat, D., and J. Hughes. 2005. Salary structure effects and the gender pay gap in academia. *Research in Higher Education* 46(6): 621–640.

Bedard, K., and D. Herman. 2008. Who goes to graduate/professional school? The importance of economic fluctuations, undergraduate field, and ability. *Economics of Education Review* 27(2): 197–210.

Bellas, M., and R. Toutkoushian. 1999. Faculty time allocations and research productivity: Gender, race and family effects. *Review of Higher Education* 22(4): 367–390.

Bender, K., and J. Heywood. 2006. Job satisfaction of the highly educated: The role of gender, academic tenure, and earnings. *Scottish Journal of Political Economy* 53(2): 253–279.

Bettinger, E., and B. Long. 2005. Do faculty serve as role models? The impact of instructor gender on female students. *American Economic Review: Papers and Proceedings* 95(2): 152–157.

Binder, M., K. Krause, J. Chermak, J. Thacher, and J. Gilroy. 2010. Same work, different pay? Evidence from a US public university. *Feminist Economics* 16(4): 105–135.

Blackaby, D., A. Booth, and J. Frank. 2005. Outside offers and the gender pay gap: Empirical evidence from the UK academic labor market. *Economic Journal* 115: F81–F107.

Blank, R. 1996. Report of the Committee on the Status of Women in the Economics Profession. *American Economic Review* 86(2): 502–506.

Blau, F., J. Currie, R. Croson, and D. Ginther. 2010. Can mentoring help female assistant professors? Interim results from a randomized trial. *American Economic Review: Papers and Proceedings* 100(2): 348–352.

Bohnet, I., A. Van Geen, and M. Bazerman. 2015. When performance trumps gender bias: Joint vs. separate evaluation. *Management Science* 62(5): 1225–1234.

Borrego, A., M. Barrios, A. Villarroya, and C. Ollé. 2010. Scientific output and impact of postdoctoral scientists: A gender perspective. *Scientometrics* 83(1): 93–101.

Boschini, A., and A. Sjogren. 2007. Is team formation gender neutral? Evidence from coauthorship patterns. *Journal of Labor Economics* 25(2): 325–365.

Bosquet, C., P. Combes, and C. Garcia-Penalosa. 2014. Gender and promotions: Evidence from academic economists in France. LIEPP Working Paper No. 29.

Bratsberg, B., J. Ragan, and J. Warren. 2010. Does raiding explain the negative returns to faculty seniority? *Economic Inquiry* 48(3): 704.

Breda, T., and M. Hillion. 2016. Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France. *Science* 353(6298): 474–478.

Broder, I. 1993. Professional achievements and gender differences among academic economists. *Economic Inquiry* 31: 116–127.

Brown, B.W., and S.A. Woodbury. 1998. Seniority, external labor markets, and faculty pay. *Quarterly Review of Economics and Finance* 38(4): 771–798.

Butler, D., and R. Butler. 2011. The Internet's effect on women's coauthoring rates and academic job market decisions: The case of political science. *Economics of Education Review* 30: 665–672.

Cainelli, G., M. Maggioni, T. Uberti, and A. de Felice. 2012. Co-authorship and productivity among Italian economists. *Applied Economics Letters* 19: 1609–1613.

Cainelli, G., M. Maggioni, T. Uberti, and A. de Felice. 2015. The strength of strong ties: How co-authorship affect productivity of academic economists? *Scientometrics* 102(1): 673.

Canes, B., and H. Rosen. 1995. Following in her footsteps? Faculty gender composition and women's choice of college majors. *Industrial and Labor Relations Review* 48(3): 486–504.

Carlin, P., M. Kidd, P. Rooney, and B. Denton. 2013. Academic wage structure by gender: The roles of peer review, performance, and market forces. *Southern Economic Journal* 80(1): 127–146.

Carlsson, F., A. Lofgren, and T. Sterner. 2012. Discrimination in scientific review: A natural field experiment on blind versus non-blind reviews. *Scandinavian Journal of Economics* 114(2): 500–519.

Carnes, M., P. Devine, L. Manwell, A. Byars-Winston, E. Fine, C. Ford, P. Forscher, C. Isaac, A. Kaatz, W. Magua, M. Palta, and J. Sheridan. 2015. Effect of an intervention to break the gender bias habit: A cluster randomized, controlled trial. *Academic Medicine* 90(2): 221–230.

Carrell, S.E., M.E. Page, and J.E. West. 2010. Sex and science: How professor gender perpetuates the gender gap. *Quarterly Journal of Economics* 125(3): 1101–1144.

Ceci, S.J., D.K. Ginther, S. Kahn, and W.M. Williams. 2014. Women in academic science: A changing landscape. *Psychological Science in the Public Interest* 15(3): 75–141.

Chen, J., Q. Liu, and S. Billger. 2012. Where do new Ph.D. Economists go? Evidence from recent initial job placements. IZA Discussion Paper No. 6990.

Chiswick, B., N. Larsen, and P. Pieper. 2010. The production of PhDs in the United States and Canada. IZA Discussion Paper No. 5367.

Conley, J., A. Onder, and B. Torgler. 2016. Are all economics graduate cohorts created equal? Gender, job openings, and research productivity. *Scientometrics* 108(2): 937–958.

Cooray, A., R. Verma, and L. Wright. 2014. Does a gender disparity exist in academic rank? Evidence from an Australian university. *Applied Economics* 46(20): 2441–2451.

Cotton, C., F. McIntyre, and J. Price. 2013. Gender differences in repeated competition: Evidence from school math contests. *Journal of Economic Behavior & Organization* 86: 52.

Danell, R., and M. Hjerm. 2013. Career prospects for female university researchers have not improved. *Scientometrics* 94(3): 999–1006.

De Paola, M., M. Ponzo, and V. Scoppa. 2015. Gender differences in attitudes towards competition: Evidence from the Italian scientific qualification. IZA Discussion Paper No. 8859.

De Paola, M., M. Ponzo, and V. Scoppa. 2016. Are men given priority for top jobs? Investigating the glass ceiling in the Italian academia. IZA Discussion Paper No. 9658.

Dee, T. 2005. A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review: Papers and Proceedings* 95(2): 158–165.

Dee, T. 2007. Teachers and the gender gaps in student achievement. *Journal of Human Resources* 42(3): 528–554.

Ding, W., S. Levin, P. Stephan, and A. Winkler. 2010. The impact of information technology on academic scientists' productivity and collaboration patterns. *Management Science* 56(9): 1439–1461.

Dobbin, F., D. Schrage, and A. Kalev. 2012. Stuck in neutral: Consequences of bureaucratic equal opportunity innovations. Working Paper, Department of Sociology, Harvard University.

Dolado, J., F. Felgueroso, and M. Alumina. 2012. Are men and women-economists evenly distributed across research fields? Some new empirical evidence. *SERIEs* 3: 367–393. doi:10.1007/s13209-011-0065-4.

Donald, S., and D. Hamermesh. 2006. What is discrimination? Gender in the american economic association. *American Economic Review* 96(4): 1283–1292.

Duch, J., X. Zeng, M. Sales-Pardo, F. Radicchi, S. Otis, T. Woodruff, and L. Amaral. 2012. The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PloS One* 7(12), E51332.

Ehrenberg, R., D. Goldhaber, and D. Brewer. 1995. Do teachers' race, gender, and ethnicity matter? Evidence from the National Educational Longitudinal Study of 1988. *Industrial and Labor Relations Review* 48(3): 547–561.

Ehrenberg, R., G. Jakubson, M. Martin, J. Main, and T. Eisenberg. 2012. Diversifying the faculty across gender lines: Do trustees and administrators matter? *Economics of Education Review* 31: 9–18.

Ellison, G., and A. Swanson. 2010. The gender gap in secondary school mathematics at high achievement levels: Evidence from the American mathematics competitions. *Journal of Economic Perspectives* 24(2): 109–128.

Else-Quest, N.M., J.S. Hyde, and M.C. Linn. 2010. Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin* 136: 103–127.

Euwals, R., and M. Ward. 2005. What matters most: Teaching or research? Empirical evidence on the remuneration of British academics. *Applied Economics* 37: 1655–1672.

Ferber, M.A., and B. Kordick. 1978. Sex differences in the earnings of Ph.D.s. *Industrial and Labor Relations Review* 31(2): 227–38.

Ginther, D. 2004. Why women earn less: Economic explanations for the gender salary gap in science. *AWIS Magazine* 33(1): 6–10.

Ginther, D., and K. Hayes. 1999. Salary and promotion differentials by gender for faculty in the humanities. *American Economic Review: Papers and Proceedings* 89(2): 397–402.

Ginther, D., and K. Hayes. 2003. Gender differences in salary and promotion for faculty in the humanities, 1977–1995. *Journal of Human Resources* 38(1): 34–73.

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.

Ginther, D., and S. Kahn. 2009. Does science promote women? Evidence from academia 1973–2001. In *Science and engineering careers in the United States*, ed. R.B. Freeman and D.F. Goroff. Chicago: University of Chicago Press for NBER.

Ginther, D., and S. Kahn. 2015. Comment on "Expectations of brilliance underlie gender distributions across academic disciplines". *Science* 349(6246): 391.

Ginther, D., S. Kahn, and W. Schaffer. 2016. Gender, race/ethnicity, and national institutes of health R01 research awards: Is there evidence of a double bind for women of color? *Academic Medicine* 91(8): 1098–1107.

Goldin, C., L. Katz, and I. Kuziemko. 2006. The homecoming of American college women: The reversal of the college gender gap. *Journal of Economic Perspectives* 20(4): 133–156.

Gordon, N., T. Morton, and I. Braden. 1974. Faculty salaries: Is there discrimination by sex, race, and discipline? *American Economic Review* 64(3): 419–427.

Griffith, A. 2014. Faculty gender in the college classroom: Does it matter for achievement and major choice? *Southern Economic Journal* 81(1): 211–231.

Groeneveld, S., K. Tijdens, and D. van Kleef. 2012. Gender differences in academic careers: Evidence for a Dutch university from personnel data 1990–2006. *Equality, Diversity and Inclusion: An International Journal* 31(7): 646–662.

Hale, G., and T. Regev. 2014. Gender ratios at top PhD programs in economics. *Economics of Education Review* 41: 55–70.

Hallock, K. 1995. Seniority and monopsony in the academic labor market: comment. *American Economic Review* 85(3): 654–657.

Hamermesh, D. 2013. Six decades of top economics publishing: Who and how? *Journal of Economic Literature* 51(1): 162–172.

Harter, C., W. Becker, and M. Watts. 2010. Time allocations and reward structures for US academic economists from 1995–2005: Evidence from three national surveys. *International Review of Economics Education* 10(2): 6–27.

Hilmer, C., and M. Hilmer. 2007. Women helping women, men helping women? Same-gender mentoring, initial job placements, and early career publishing success for economics PhDs. *American Economic Review* 97(2): 422–426.

Hilmer, C., and M. Hilmer. 2010. Are there gender differences in the job mobility patterns of academic economists? *American Economic Review* 100(2): 353–357.

Hilmer, C., M. Hilmer, and M. Ransom. 2012. Fame and the fortune of academic economists: How the market rewards influential research in economics. IZA Discussion Paper No. 6960.

Hoffman, F., and P. Oreopoulos. 2009. A professor like me. The influence of instructor gender on college achievement. *Journal of Human Resources* 44(2): 479–494.

Hyde, J.S., and J. Mertz. 2009. Gender, culture, and mathematics performance. *Proceedings of the National Academy of Science* 106: 8801–8809.

Joecks, J., K. Pull, and U. Backes-Gellner. 2014. Childbearing and (female) research productivity: A personnel economics perspective on the leaky pipeline. *Journal of Business Economics* 84: 517–530.

Kahn, S. 1993. Gender differences in academic career paths of economists. *American Economic Review: Papers and Proceedings* 83(2): 52–56.

Kahn, S. 1995. Women in the economics profession. *Journal of Economic Perspectives* 9(4): 193–205.

Kahn, S. 2012. Gender differences in academic promotion and mobility at a major Australian university. *Economic Record* 88: 407.

Kahn, S., and D. K. Ginther. 2016. Accounting for the gender pay gap in STEM salaries. Mimeo/Boston University.

Kalev, A., F. Dobbin, and E. Kelly. 2006. Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American Sociological Review* 71: 589–617.

Kaszubowski, M., and J. Wolszczak-Derlacz. 2014. Salary and reservation wage gender gaps in Polish academia. GUT Faculty of Management and Economics Working Paper Series A, No. 1 (19).

Krapf, M., H. Ursprung, and C. Zimmermann. 2014. Parenthood and productivity of highly skilled labor: Evidence from the groves of academe. IZA Discussion Paper No. 7904.

Krause, A., U. Rinne, and K. Zimmermann. 2012. Anonymous job applications of fresh Ph.D. Economists. *Economics Letters* 117(2): 441–444.

Landaud, F., S.-T. Ly, and E. Maurin. 2016. Competitive schools and the gender gap in the choice of field of study. CEPR Discussion Paper No. DP11411. Available at SSRN: http://ssrn.com/abstract=2814086.

Lavy, V., and E. Sand. 2015. On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases. NBER Working Paper 20909.

Leslie, S., A. Cimpian, M. Meyer, and E. Freeland. 2015. Expectations of brilliance underlie gender distributions across academic disciplines. *Science* 347(6219): 262.

Lewis, M., and K. McGoldrick. 2001. Moving beyond the masculine neoclassical classroom. *Feminist Economics* 7(2): 91–103.

Ley, T., and B. Hamilton. 2008. The gender gap in NIH grant applications. *Science* 322(5907): 1472–1474.

Lissoni, F., J. Mairesse, F. Montobbio, and M. Pezzoni. 2011. Scientific productivity and academic promotion: A study on French and Italian physicists. *Industrial and Corporate Change* 20(1): 253–294.

Manchester, C., and D. Barbezat. 2013. The effect of time use in explaining male–female productivity differences among economists. *Industrial Relations: A Journal of Economy and Society* 52(1): 53–77.

Manchester, C., L. Leslie, and A. Kramer. 2010. Stop the clock policies and career success in academia. *American Economic Review: Papers & Proceedings* 100: 219–223.

Manchester, C., L. Leslie, and A. Kramer. 2013. Is the clock still ticking? An evaluation of the consequences of stopping the tenure clock. *ILR Review* 66(1): 3–31.

Maske, K., G. Durden, and P. Gaynor. 2003. Determinants of scholarly productivity among male and female economists. *Economic Inquiry* 41(4): 555–564.

McDowell, J.M., and K. Smith. 1992. The effect of gender sorting on propensity to coauthor: Implications for academic promotion. *Economic Inquiry* 30(1): 68–82.

McDowell, J., L. Singell Jr., and J. Ziliak. 1999. Cracks in the glass ceiling: Gender and promotion in the economics profession. *American Economic Review: Papers & Proceedings* 89(2): 397–402.

McDowell, J., L. Singell Jr., and J. Ziliak. 2001. Gender and promotion in the economics profession. *Industrial and Labor Relations Review* 54(2): 224–244.

McDowell, J., L. Singell Jr., and M. Slater. 2006. Two to tango? Gender differences in the joint decision to publish and coauthor. *Economic Inquiry* 44(1): 153–168.

McGoldrick, K., and P. Schuhmann. 2002. Instructor gender and student registration: An analysis of preferences. *Education Economics* 10(3): 241–260.

Mixon, F., and L. Trevino. 2005. Is there gender discrimination in named professorships? An econometric analysis of economics departments in the US South. *Applied Economics* 37: 849–854.

Moss-Racusin, C., J. Dovidio, V. Brescoll, M. Graham, and J. Handelsman. 2012. Science faculty's subtle gender biases favor male students. *PNAS* 109(41): 16474–16479.

National Research Council 2010. *Gender differences at critical transitions in the careers of science, engineering, and mathematics faculty.* Committee on Gender Differences in the Careers of Science, Engineering, and Mathematics Faculty; Committee on Women in Science, Engineering, and Medicine; Committee on National Statistics; Policy and Global Affairs; Division of Behavioral and Social Sciences and Education; National Research Council.

Niederle, M., and L. Vesterlund. 2010. Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives* 24(2): 129–144.

Penner, A.M. 2008. Gender differences in extreme mathematical achievement: An international perspective on biological and social factors. *American Journal of Sociology* 114: 138–170.

Pope, D., and J. Sydnor. 2010. Geographic variation in the gender differences in test scores. *Journal of Economic Perspectives* 24(2): 95–108.

Ransom, M. 1993. Seniority and monopsony in the academic labor market. *American Economic Review* 83(1): 221.

Rask, K., and E. Bailey. 2002. Are faculty role models? Evidence from major choice in an undergraduate institution. *Journal of Economic Education* 33(2): 99–124.

Robinson, M., and J. Monks. 1999. Gender differences in earnings among economics and business faculty. *Economics Letters* 63(1): 119–125.

Sabatier, M. 2010. Do female researchers face a glass ceiling in France? A hazard model of promotions. *Applied Economics* 42(16): 2053–2062.

Schulze, G., S. Warning, and C. Wiermann. 2008. What and how long does it take to get tenure? The case of economics and business administration in Austria, Germany and Switzerland. *German Economic Review* 9(4): 473–505.

Sege, R., N. Nykiel-Bub, and S. Selk. 2015. Sex difference in institutional support for junior biomedical researchers. *Journal of the American Medical Association* 314(11): 1175–1177.

Takahashi, A., and S. Takahashi. 2015. Gender promotion differences in economics departments in Japan: A duration analysis. *Journal of Asian Economics* 41: 1–19.

Taylor, S., B. Fender, and K. Burke. 2006. Unraveling the academic productivity of economists: The opportunity

costs of teaching and service. *Southern Economic Journal* 72(4): 846–859.
Toumanoff, P. 2005. The effects of gender on salary-at-hire in the academic labor market. *Economics of Education Review* 24: 179–188.
Toutkoushian, R., M. Bellas, and J. Moore. 2007. The interaction effects of gender, race, and marital status on faculty salaries. *Journal of Higher Education* 78(5): 572–601.
Wai, J., M. Cacchio, M. Putallaz, and M. Makel. 2010. Sex differences in the right tail of cognitive abilities: A thirty year examination. *Intelligence* 38: 412–423. doi:10.1016/j.intell.2010.04.006.
Ward, M. 2001a. The gender salary gap in British academia. *Applied Economics* 33(13): 1669–1681.
Ward, M. 2001b. Gender and promotion in the academic profession. *Scottish Journal of Political Economy* 48(3): 283–302.
Ward, M., and P. Sloane. 2000. Non-pecuniary advantages versus pecuniary disadvantages; Job satisfaction among male and female academics in Scottish universities. *Scottish Journal of Political Economy* 47(3): 273.
Warman, C., F. Woolley, and C. Worswick. 2010. The evolution of male–female earnings differentials in Canadian universities, 1970–2001. *Canadian Journal of Economics-Revue Canadienne D'Economique* 43(1): 347–372.
Williams, W., and S. Ceci. 2015. National hiring experiment reveal 2:1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences* 112(17): 5360–5365.
Wolfinger, N.H., M.A. Mason, and M. Goulden. 2008. Problems in the pipeline: Gender, marriage, fertility and the ivory tower. *Journal of Higher Education* 79(4): 388–405.
Zinovyeva, N., and, M. Bagues. 2011. Does gender matter for academic promotion? Evidence from a randomized natural experiment. IDEAS Working Paper Series from RePEc.

Gender Differences (Experimental Evidence)

Catherine C. Eckel

Abstract

Laboratory experiments find differences between women and men in three main areas: altruism, risk aversion and competition. The types of experiments and findings are described, and findings summarized. These results parallel similar findings in other social sciences, and are consistent with observed differences in the field.

Keywords

Altruism; Altruism in experiments; Charitable giving; Competition; Cooperation; Dictator game; Experimental economics; Gender differences; Gender gap; Overconfidence; Preference elicitation; Risk aversion; Social dilemma; Ultimatum game; Women's work and wages

JEL Classifications

C9

Henry Higgins famously enquires, 'Why can't a woman be more like a man?' While others might phrase it differently, the question of how and why women and men differ is the subject of hundreds of books and articles every year. Experimental economists have investigated gender differences in at least three areas: cooperation and altruism, attitudes toward risk, and preferences for engaging in competitive activities. While this research has proceeded largely independently of other social sciences, the results across fields are parallel.

Most experimental research on gender differences is motivated by an interest in the persistent gender gap in earnings, with women earning significantly less than men even after adjusting for productivity related differences in education, experience, choice of employment, and so on (Weichselbaumer and Winter-Ebmer 2005). While this gender gap has diminished since the 1970s it has not disappeared. Attention is also directed at the fact that women are underrepresented in leadership positions. Within economics, the Committee on the Status of Women in the Economics Profession of the American Economic Association keeps tabs on the progress of women. Their most recent survey shows that women continue to lag behind men in their progress towards higher academic ranks. While women earned 20%

of Ph.D.s in economics in the 1980s and 25% in the 1990s, only 8.3% of full professors in Ph.D. granting departments are women (see Committee on the Status of Women 2007).

Differences in the behaviour of women and men are extensively documented in research in psychology and sociology. An overview of this work, which covers differences in ability, personality, leadership styles, aggression, competitiveness and so on, can be found in Rhoads (2004) and Maccoby (1998) among many others. Experiments have examined gender differences in situations involving salient monetary incentives since Rapoport and Chammah (1965), who explored variations of the Prisoner's Dilemma (PD) game. Early experimental work in psychology and sociology tended to involve this game, or related social dilemma (SD) games, with mixed results. In games with this incentive structure – where each player has a dominant strategy to free ride, but group payoffs are maximized by choosing a cooperative strategy – many studies have found that women are more cooperative, and many that they are less so.

Experimental research in economics focuses on examining the types of preferences that might be related to the gender gap: those that relate to cooperating, taking risks and competing. Compared to the stereotypical male person, the stereotypical female person is more altruistic and cooperative, and more averse to risk and competition. Partial surveys of research in experimental economics on gender differences are provided by Eckel and Grossman (2008a), which focuses on altruism and cooperation, and (2008b), which surveys studies of risk aversion, and a more comprehensive review is contained in Croson and Gneezy (2007).

Cooperation

If there is no systematic difference between the sexes in their play of PD and SD games, can we abandon this element of the stereotype and conclude that women are no more cooperative than men when money is at stake? Eckel and Grossman (1998) were the first to point out that these games confound two possible differences in the preferences of women and men. Suppose that, true to stereotype, women are both more altruistic and more risk averse. Altruistic preferences imply that women will be more likely to choose a cooperative strategy in PD and SD games. However, risk aversion implies just the opposite. The cooperative strategy is also the risky strategy; a cooperator risks being exploited, with corresponding low earnings. The best choice for an altruistic, risk-averse person would depend on the parameters of the game, that is, the trade-off between the gain to cooperation and the penalty if one is betrayed. Thus the games that have been most commonly used to measure cooperation may be confounded by risk aversion.

Eckel and Grossman's (1998) strategy was to separate altruism from risk aversion. In a double-anonymous dictator game, where there is no financial (or social) risk, they report that women give about twice as much to an anonymous partner. This result has not always been replicated by subsequent studies, and behaviour can vary with the characteristics of the recipients when they are known, but overall it is rare to find a situation where men are more altruistic. In more complex experiments (Andreoni and Vesterlund 2001; Dickinson and Tiefenthaler 2002), subjects make a series of dictator decisions in tokens, where the tokens have different exchange rates for each of the players. In these games men tend to maximize efficiency, allocating more to the partner with the better exchange rate, while women tend to try to equalize earnings. Thus men appear more altruistic at exchange rates that benefit the recipient. At equal exchange rates, women give more than men. Moreover, in studies where subjects can give to a charitable organization, Eckel and Grossman consistently find that women give more than men (for example, Eckel and Grossman 2003).

Another experimental environment that has received a great deal of attention is the ultimatum game, which suffers from a similar problem. In that game, a person might make a generous offer because of altruism or because of risk aversion; similarly, a person might accept a low offer for multiple reasons. The greater altruism and risk aversion attributed to women implies more

generous ultimatum offers by women. However, results in this game are mixed (Eckel and Grossman 2001; Solnick 2001). Women and men make similar offers on average, but, more importantly, both make lower offers to women than to men, suggesting a commonly held belief that women will accept lower offers (because they are more altruistic?). On the respondent side, the results of these two studies are contradictory. In general, however, results indicate that women are more likely to accept an offer of a given size than the reverse.

A higher degree of altruism is consistent with lower wages, with more altruistic persons both requesting and accepting lower wage offers. It is worth pointing out that (to my knowledge) no studies have tested the external validity of these measures of altruism; that is, economists have little or no knowledge of how well laboratory-elicited preferences 'predict' how people behave as they go about their daily lives. While lab decisions are real in the sense that there are resource consequences to decisions, the context is very different from field decisions. However, there are several studies that explicitly examine cultural context, and find a positive relationship between how groups play a public goods game and how they harvest natural resources (Carpenter and Seki 2006). There is also some evidence that the gender gap in earnings is smaller in the nonprofit sector, where altruistic preferences might be especially valuable (Leete 2000).

Risk Aversion

Like cooperation, gender differences in risk aversion have been much studied in fields outside economics. In most situations, greater risk taking by men is well documented (Byrnes et al. 1999). Economists have a rather narrow way of thinking about risk aversion compared with other social scientists; we view preferences as represented by a utility function that evaluates alternatives across all decision-making domains. Diminishing marginal utility of income or wealth produces risk aversion: the expected value of a gamble always has higher utility than the gamble itself. (Of course constant marginal utility implies risk neutrality, and increasing marginal utility risk seeking.) This view of risk aversion implies that any task that measures the curvature of the utility function in money should give a good measure of risk attitudes that is then applicable across all situations. Experimental economists have developed different games that do just that.

Like their counterparts in the other social sciences, economists tend to find women more risk averse than men, though both are surprisingly risk averse considering the level of stakes in our games. Though the difference is not always statistically significant, it is rare that it goes the other way. However, there is a potential problem with commonly used measures that might distort the gender difference. The experiments tend to be complicated, requiring a relatively high level of mathematical ability to be clearly understood. This is not a big problem if any resulting 'noise' does not bias the measure. Unfortunately, there is some indication that difficult tasks cause low-ability subjects to make systematically different choices. To the extent that mathematical ability is correlated with gender, this could bias inferences about differences in risk aversion.

A popular initial experiment used a risky version of the Becker et al. (1964) (BDM) preference elicitation procedure. This mechanism elicits subjects' valuations for gambles with various probabilites of winning a particular prize. To make it incentive compatible, this mechanism requires two stages. In the first the subject writes down a minimum selling price for a gamble that pays off X dollars with probability p. In the second a random price is drawn from a uniform distribution between 0 and X dollars. If the drawn price is above the elicited price, the subject sells the gamble, and if not the subject plays the gamble. This mechanism has been criticized for its complexity, and for the low incentive for accuracy with low-probability gambles. Subjects with low maths ability may be more likely to overvalue these low probability gambles because they are confused by the second stage of the game. For example, consider a 0.05 probability of winning 10 dollars; any value drawn in the second stage is likely to be far above the expected value of the

gamble, so it may not be worthwhile for the subject to bother calculating his reservation value, and he is likely to err on the high side. This would tend to make subjects look less risk averse. Indeed, BDM studies tend to find fewer risk-averse and more risk-seeking subjects.

A cleverly designed game developed by Holt and Laury (2002) has subjects choose between pairs of lotteries that are constructed so as to easily 'back out' a coefficient of relative risk aversion for a specified utility functional form. This game is easily comprehended by college students, and produces intuitively appealing results in educated populations (Andersen et al. 2006). However, there is some evidence it is less successful for less literate populations, limiting its usefulness in the field (Dave et al. 2007). Like the BDM procedure, failure to account for differences in mathematical ability may distort estimates of gender differences.

A third type of game involves fewer choices among simpler, 50/50 gambles (Binswanger 1980; Eckel and Grossman 2002). A subject chooses her favourite from among a set of 50–50 gambles that vary in risk and expected return. The experiment allows categorization of subjects into ordered categories, from most to least risk averse. There is some evidence that this experiment is easier to comprehend for populations with low mathematical literacy, although the trade-off is that the measure is coarser than the others described above. One troubling result

even for educated groups is that different measures of risk aversion completed by the same set of people tend to exhibit low correlations across measures, suggesting that our underlying construct may need some work.

The experiments above all involve individual decisions. Additional indirect evidence of risk aversion can be found in a market environment. Women are more likely than men to overbid in first-price auctions, behaviour that can be caused by risk aversion. Chen et al. (2005) find that women tend to overbid, but women's bids are most like men's when oestrogen levels are lowest, suggesting a biological mechanism driving greater risk aversion.

Gender differences are typically, but not always, found across all experiments designed to measure risk aversion. Women are more risk averse across environments. Several studies have begun to examine external validity of the measures. In general, lab measures of risk attitudes have low (though sometimes statistically significant) correlations with decisions in other lab experiments, and low correlations with risky field behaviours, such as buying an extended warranty for an automobile or computer (Moore 2002). Risk attitudes also are related to a person's willingness to borrow to finance higher education expenditure (Eckel et al. 2007). Many current studies will further examine external validity of experimental preference measures. As with altruism, to my knowledge, no study has related experimental risk measures to employment earnings. However, field studies tend to confirm gender stereotypes, with women investing in more conservative portfolios (Sunden and Surette 1998), more likely to buy warranties (Moore 2002), and more likely to negotiate contracts with larger salary components and smaller performance-related components (Chauvin and Ash 1994). The outstanding question is whether experimentalists can measure risk attitudes in experimental games in a way that meaningfully predicts risky choices in field and, in particular, employment settings.

Competition

Women do not like competition. Psychologists have long known that girls are less competitive than boys, that they play different games and avoid competitive situations. For example, Maccoby (1998) quotes many such studies, including one showing that, in same-sex groups of fourth and sixth graders, boys spontaneously engaged in competitive activities 50% of the time, while girls engaged in such play only one per cent of the time (1998, p. 39). Men do not merely like competition, they also do better when a situation is more competitive. Rhoads (2004) surveys work in this area and gives dozens of examples. Some authors have used these differences to argue that women are inherently ill-suited to the workplace (Browne 2002), and others that women have an advantage because competition does not get in the

way of making the best decisions (Helgesen 1990). The taste for competition is no doubt related to men's higher levels of confidence; overconfidence can also interfere with profit-maximization, as Barber and Odean (2001) show in a study of online stock trading.

Experimental economists have discovered this, too: Gneezy et al. (2003) show in a lab experiment that introducing competition makes men, but not women, more productive in solving mazes. The study compares work performance under two types of compensation: piece rates, where workers are paid by the maze, and winner-take all, tournament rate, where only the highest producer is paid. Women work about the same under the two schemes, while men work significantly harder for the tournament payment. This result spurred two additional studies where women and men choose their preferred compensation rate. In the first, Gupta et al. (2005) again use mazes and find that 60% of men and 34% of women choose the tournament rate. Niederle and Vesterlund (2007) are careful to choose a task where women and men perform the same under piece and tournament rates – solving easy maths problems. Here again, men are more likely to choose the tournament (73% compared to 35%). This effect remains after controlling for subjects' measured ability as well as their own perceptions of their abilities; thus the result is not due to overconfidence. Men sacrifice earnings in this game because low-ability men choose the tournament, but women lose more and so earn less than men because high-ability women shy away from the tournament.

If women avoid competition, this, too, may have consequences for earnings. If a preference to avoid competition transfers from the lab to the field, then it is likely to affect the earnings of women. As with cooperation and risk, more study is needed to verify the external validity of the lab-based measures of aversion to competition.

Conclusion

Laboratory experiments show a collection of preferences that differ, on average, between the sexes. Women tend to be more altruistic, risk averse, and competition averse. This pattern of preferences could lead to patterns of behaviour that result in lower wages for women, such as accepting low offers or avoiding competitive situations. For example, Babcock and Laschever (2003) find that lower average starting salaries for women public policy graduates are the result of differences in the way men and women treat job offers. Women tend to accept the best offer they receive from potential employers; men, by contrast, respond to an offer by asking for more. This behaviour seems very much like that observed in the lab, and suggests that altruism, risk aversion, and competition aversion may play a role in explaining this.

The results of economics lab experiments are largely consistent with research from the other social sciences, and psychology in particular. Economics experiments are conducted in settings where payoffs are salient, and where there is no deception. This work has not only confirmed but also legitimized research on gender differences for economists. The importance of the work is to show that individual differences in preferences, whether by nature or nurture, can be substantial, and are correlated with observable characteristics of individuals. Decision-making in the workplace occurs in a much more complex environment, making it difficult or impossible to sort out the effects of the various dimensions of preferences. However, lab experiments allow a much higher degree of control over the environment so that variability in specific aspects of preferences can be isolated.

See Also

- ▶ Gender Roles and Division of Labour
- ▶ Social Preferences
- ▶ Women's Work and Wages

Bibliography

Andersen, S., G.W. Harrison, M.I. Lau, and E.E. Rutstrom. 2006. *Eliciting risk and time preferences*. Working Paper 05–24, Department of Economics, College of Business Administration, University of Central Florida.

Andreoni, J., and L. Vesterlund. 2001. Which is the fair sex? Gender differences in altruism. *Quarterly Journal of Economics* 116: 293–312.

Babcock, L., and S. Laschever. 2003. *Women don't ask: Negotiation and the gender divide*. Princeton: Princeton University Press.

Barber, B.M., and T. Odean. 2001. Boys will be boys: Gender, overconfidence and common stock investing. *Quarterly Journal of Economics* 116: 261–292.

Becker, G.M., M.H. DeGroot, and J. Marschak. 1964. Measuring utility by a single-response sequential method. *Behavioral Science* 9: 226–232.

Binswanger, H.P. 1980. Attitudes toward risk: Experimental measurement in rural India. *American Journal of Agricultural Economics* 62: 395–407.

Browne, K. 2002. *Biology at work: rethinking sexual equality*. New Brunswick: Rutgers University Press.

Byrnes, J., D.C. Miller, and W.D. Schafer. 1999. Gender differences in risk taking: A meta analysis. *Psychological Bulletin* 125: 367–383.

Carpenter, J., and E. Seki. 2006. Competitive work environments and social preferences: Field experimental evidence from a Japanese fishing community. Contributions to *Economic Analysis & Policy* 5, Article 2. Abstract online. Available at http://www.bepress.com/bejeap/contributions/vol5/iss2/art2. Accessed 16 June 2007.

Chauvin, K.W., and R.A. Ash. 1994. Gender earnings differentials in total pay, base pay, and contingent pay. *Industrial and Labor Relations Review* 47: 634–649.

Chen, Y., P. Katuscak, and E. Ozdenoren. 2005. Why can't a woman bid more like a man? Working paper, University of Michigan.

Committee on the Status of Women. 2007. Winter newsletter. Online. Available at http://www.cswep.org/newsletters/CSWEP_nsltr_Winter2007.pdf. Accessed 12 June 2007.

Croson, R., and U. Gneezy. 2007. Gender differences in preferences. Working paper, University of California, San Diego.

Dave, C., C.C. Eckel, C.A. Johnson, and C. Rojas. 2007. Eliciting risk preferences: Heterogeneity, stability and precision. Working paper, University of Texas at Dallas.

Dickinson, D., and J. Tiefenthaler. 2002. What is fair? Experimental evidence. *Southern Economic Journal* 69: 414–428.

Eckel, C.C., and P.J. Grossman. 1998. Are women less selfish than men? Evidence from dictator games. *Economic Journal* 108: 726–735.

Eckel, C.C., and P.J. Grossman. 2001. Chivalry and solidarity in ultimatum games. *Economic Inquiry* 39: 171–188.

Eckel, C.C., and P.J. Grossman. 2002. Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior* 23: 281–295.

Eckel, C.C., and P.J. Grossman. 2003. Rebates versus matching: Does how we subsidize charitable contributions matter? *Journal of Public Economics* 87: 681–701.

Eckel, C.C., and P.J. Grossman. 2008a. Differences in the economic decisions of men and women: Experimental evidence. In *Handbook of experimental economics results*, ed. C. Plott and V.L. Smith. New York: Elsevier forthcoming.

Eckel, C.C., and P.J. Grossman. 2008b. Men, women and risk aversion: Experimental evidence. In *Handbook of experimental economics results*, ed. C. Plott and V.L. Smith. New York: Elsevier forthcoming.

Eckel, C.C., C.A. Johnson, C. Montmarquette, and C. Rojas. 2007. Debt aversion and the demand for loans for postsecondary education. *Public Finance Review* 35: 233–262.

Gneezy, U., M. Niederle, and A. Ructichini. 2003. Performance in competitive environments: Gender differences. *Quarterly Journal of Economics* 118: 1049–1074.

Gupta, N.D., A. Poulsen, and M.C. Villeval. 2005. *Do (wo) men prefer (non) competitive jobs?* Working paper, Institute for the Study of Lavoe (IZA).

Helgesen, S. 1990. *The female advantage: Women's ways of leadership*. New York: Doubleday.

Holt, C.A., and S.K. Laury. 2002. Risk aversion and incentive effects. *American Economic Review* 92: 1644–1655.

Leete, L. 2000. Wage equity and employee motivation in nonprofit and for-profit organizations. *Journal of Economic Behavior & Organization* 43: 423–446.

Maccoby, E. 1998. *The two sexes*. Cambridge: Harvard University Press.

Moore, E. 2002. *An investigation into the demand for service contracts*. Ph.D. thesis, Department of Economics, Virginia Polytechnic Institute and State University.

Niederle, M., and L. Vesterlund. 2007. Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics* 122(3), forthcoming.

Rapoport, A., and A.M. Chammah. 1965. Sex differences in factors contributing to the level of cooperation in the prisoner's dilemma game. *Journal of Personality and Social Psychology* 2: 831–838.

Rhoads, S.E. 2004. *Taking sex differences seriously*. San Francisco: Encounter Books.

Solnick, S. 2001. Gender differences in the ultimatum game. *Economic Inquiry* 39: 189–200.

Sunden, A.E., and B.J. Surette. 1998. Gender differences in the allocation of assets in retirement savings plans. *American Economic Review* 88: 207–211.

Weichselbaumer, D., and R. Winter-Ebmer. 2005. A metaanalysis of the international gender wage gap. *Journal of Economic Surveys* 19: 479–511.

Gender Roles and Division of Labour

Joyce P. Jacobsen

Abstract

All human societies exhibit some degree of division of labour by gender. These divisions continue to exist as participation in paid work has increased over time. Gender divisions occur between household tasks, between unpaid and paid work, and within paid work. Economists have explained these divisions through reliance on essentialist arguments and/or the fundamental economic concepts of efficiency of specialization and division of labour, and investment in human capital. However, gender discrimination can also cause division of labour, and the feedback effects of such discrimination make it difficult to untangle the causes of the gender division of labour.

Keywords

Affirmative action; Becker, G.; Capitalism; Gender roles and division of labour; Household economy; Human capital; Intrahousehold welfare; Labour market discrimination; Marriage markets; Non-market work; Occupational segregation; Patriarchy; Social norms; Socialism; Technical change; Women's work and wages; Work–leisure trade-off

JEL Classifications

J16

All human societies have a gender-related division of labour, although the particulars of division vary across time and culture. It is generally agreed to be a pre-capitalist phenomenon, based on anthropological and historical information, and related to the widespread existence of patriarchy, that is, male-controlled and male-favouring social systems. Let us start by considering the arguments for why there would be gender-related task specialization in a non-market setting, that is, in a pre-capitalist society.

Gender Roles and Task Specialization

Most of the arguments posited for why a gender-related division of labour exists have been essentialist. The general analytical approach has been to posit a specific biological sex-related difference and then show that it leads to gender task specialization. The factor can be a sex difference in ability, including in some models the ability to develop or learn particular types of skills – and thus specialization leads to efficiency in production. Or the determining factor can be a sex difference in preference or taste – and thus specialization leads directly to utility maximization. In the first category, analysts have posited differences in childbirth and child-raising abilities, fecundity, physical strength, aggression/dominance/coalition-building/risk-taking, or cognitive differences (Fausto-Sterling 1985; Duley et al. 1986; Becker 1991; Siow 1998). Gender differences in ability need not imply unbiased gains for both sexes. Indeed, if patriarchy arises because of one or more of these differences, then it is also plausible that the gender division of labour favours men, or at least some men. In the second category, analysts have posited such factors as sex differences in preferences for spouse's age (Elul et al. 2002), differences in caring for children and others (Folbre 1995), and different preferences for meaningful work and other job characteristics over money (Brown and Corcoran 1997).

A smaller set of analysts has presented non-essentialist arguments, where the general approach has been to posit that a division of labour is efficient and that some specialized human capital must be acquired initially in order to improve efficiency further; in addition, human capital in the form of specialized experience can be developed through continued application to the specific task. Therefore, in order to maximize output, societies should train some people to do

one type of task, and others to do something different, and leave them in those roles for extended periods of time. Formal models of this type then link the labour market with the marriage market to consider the coordination problem and societal output maximization in arguing why it is important that people of one type, for example, sex, be assigned to one sort of task (Hadfield 1999; Engineer and Welling 1999; Baker and Jacobsen 2006). If some types of output can be produced and traded only with the household, then it is important to match people of different types within marriages. Thus men should do one type of task and women another to reduce the coordination problem. The dilemma for these models is to explain why particular tasks are assigned to men and others assigned to women, particularly if some tasks are preferable on some dimension (for example, they have higher prestige or portability) and also tend empirically to be assigned to men. Thus these models often have to fall back on an essentialist starting point in order to determine initial assignment. They can then argue that societal dynamics in determining future gender assignments are affected by the initial assignment and by technological change.

The non-essentialist arguments do a better job than the essentialist arguments of explaining why societies have prescribed gender roles rather than allowing for flexibility of task assignment based on actual individual abilities. A strong essentialist argument would broach no conflict between biological sex and gender roles, yet we see deviations from gender roles by individuals, weakening the essentialist argument. Thus essentialists need a more nuanced approach in which biological sex (whether chromosonal or hormonal) leads to different probabilities of particular outcomes, or different distributions of traits. Then an argument based on efficiency of division of labour, along with the need to make specialized human capital investments early on in children's lives (Becker 1991), leads society to assign gender roles based on average or modal outcomes by sex.

The question then arises of how to deal with deviations from gender roles. People can generally articulate gender norms, that is, roles that are considered sex-appropriate, and know when they or others are violating them. Gender roles can relate also to age, and also may have caste or class or racial/ethnic aspects, so tasks may be assigned differentially based on these other dimensions, too. Societies deal with deviations in different ways, including complete proscription; allowing people to change later on at an efficiency loss; allowing for exceptions only if people show particular deviant traits early on; and laissez-faire. Akerlof and Kranton (2000) posit that gender identity appears in the utility function and that deviations from sex-appropriate gender identity cause utility loss for individuals. Badgett and Folbre (2003) discuss the potential penalty that one may face in the marriage market for being in a gender-non-conforming occupation. Changes in social norms regarding the social construction of sexuality may have had some effect in reducing these losses (Matthaei 1995). Deviation tends to appear only within a dualist system; indeed, the prominence of gender duality in most, but not all, cultures is notable. In cultures with a third gender role, such persons either are assigned to specific reserved occupations or must conform to the cross-gender role if sex is not aligned with gender (Jacobsen 2006).

Gender Division of Labour Between Non-market and Market Work

With the development of markets for paid labour, division additionally occurs between unpaid (non-market) and paid (market) work. It is a general observation across societies that women are more likely to specialize in non-market work and men in market work, or women to divide time between non-market and market work, and men to specialize in market work. As a first step in explaining this pattern, the neoclassical approach posits that division of labour is efficient. The household is considered the nexus for production and consumption of non-market commodities. Thus division of labour within the household occurs, with some members specializing in market work, others in non-market work. In models of the modern household, children are generally

treated as consumption goods, or sometimes as investments, when previously they were thought of as additional suppliers of market or non-market labour. In order to motivate the particular gender division of labour, writers fall back generally on one or more essentialist arguments for why women do non-market work, in particular the relation to bearing and raising children. If the division of labour between non-market and market falls along gender lines, the marriage market may then be conceptualized as the market for non-market, or spousal, labour (Grossbard-Shechtman 1993).

While the argument that specialization and division of labour is more efficient might hold in a static framework, it is not obvious that this is necessarily optimal in a dynamic framework, at least not for both parties. Specialization in non-market labour is the less desirable specialty as it limits the market for one's services by definition; indeed, if there is no marriage, there is no market for one's services at all. Thus models have explicitly linked the household division of labour to the operation of the marriage market, whether these concerns are taken into account in settling on a division of output before entering into marriage (Lundberg and Pollak 1994) or within marriages through ongoing negotiation over distribution of the household's product.

It is also problematic to argue that specialization decisions are made solely on the basis of relative productivity. If there is discrimination in the labour market such that women are paid below their actual marginal product, then women's comparative advantage is more likely to lie in household work. Lower wages can also lead to intermittent labour force attachment, which leads back to lower wages (Gronau 1988). Becker (1991) argues that effort as well as time must be allocated across types of labour; if women expend more effort on household work, then they have less effort to exert on market work, thus receiving lower wages. Also, if women train to do household work, whether or not they have a greater aptitude for it, then they are more likely to have comparative advantage in household work (and the opposite for men and market work). Thus the efficiency arguments can be self-fulfilling.

One way to attempt to untangle these feedback mechanisms is to see what happens when society experiences changes through technology, political upheaval, or other factors. But the processes of political and economic change have had mixed effects on the gender division of labour, even as they have had large effects in changing the nature of work and the mix between household and market work. Some hold the view that capitalism accentuates the gender division of labour through accentuating the division between household and market work, while others think capitalism is useful in reducing the gender division of labour. Some argue that patriarchy and capitalism are mutually reinforcing (Hartmann 1976; Humphries 1991) and that socialism needs to include the overturning of both systems (Engels 1884) including reassignment or eradication of patriarchy-enforcing property rights (Braunstein and Folbre 2001). Cases of transition from capitalism to socialism – and for some countries back again – have provided mixed evidence; in practice, socialism appears to have increased women's total work time, increasing their paid work without decreasing their unpaid work (Jacobsen 2006).

In practice, most people in modern societies do both paid and unpaid work, whether in a given time period or across the life cycle. Technological change affects gender work assignments over time whenever it is non-neutral with respect to initially gendered task assignments. It appears that the particular form in which technological change has occurred has made capital complementary to women's market work (Galor and Weil 1996) and substitutable for women's non-market work (Greenwood et al. 2005). Real wages have been rising for women over the past century. The net effect has been to reduce women's time spent in non-market work and to increase their time spent in market work.

Gender Segregation in Market Work

Even as paid labour has become more extensive and women have increasingly participated in paid work, extensive gender segregation persists

across time and space in labour markets (Anker 1998; Jacobsen 2006). Market work for women often still emulates their areas of traditional female-dominated non-market work, such as child care and teaching of young children, nursing and eldercare, and food preparation and service. Men still dominate the occupations that have required more physical strength, and in industrialized societies are more likely to work in outdoor occupations. What is harder to explain along essentialist or traditionalist lines is why there would be gender segregation for other types of occupations that have arisen later in economic development, such as various types of professions.

Economists have advanced various explanations for occupational gender segregation. Again, many rely on essentialist arguments regarding differences in abilities and/or preferences to explain why women and men would choose different paid work. One approach is to argue that some jobs are more compatible with non-market duties. Thus, if women are doing most of the non-market work (which begs the question of why they are doing it), they must choose jobs that allow for this balance, including those that allow for part-time work. A dynamic version of this argument is that women know they will be balancing paid work with non-market work, particularly during their childbearing years, and thus choose occupations that are potentially more compatible with this lower level of attachment to paid work. Notably, even in occupations where women have increased their representation substantially, there is within-occupation gender segregation along various lines, including sub-specialties, firm size, employer type (for-profit, non-profit, government), and so on. These patterns appear in many cases to be consistent with an argument that women prefer more flexible employment, that is, jobs that require less travel and less overtime work, and allow for part-time and/or flexible hours.

To switch from a supply-side to a demand-side focus, other economists have argued that gender segregation is driven by employers' choosing whom to hire, not by employees' choosing where to work. Re Becker (1971), employers may get utility directly from discriminating or simply from maintaining social norms. Employer discrimination can occur if there is insufficient competition from non-discriminating employers to drive out discriminating employers. Segregation can occur without loss of profits in Becker's employee and customer discrimination models. Male-dominated unions and other professional organizations can keep women out of particular occupations by denying them training (Fawcett 1892). Statistical discrimination is another potential explanation of gender segregation, with the usual chicken-and-egg problem: employers don't see women doing the non-traditional task, so cannot tell whether women are good at it (Lundberg and Startz 1983). Some have also called into question the implicit assumption of exogeneity of gender-linked preferences if pre-labour market treatment of girls and boys is different (Corcoran and Courant 1987).

There is interesting evidence regarding the instability of gender integration from studying workers like clerks, bank tellers, and schoolteachers, whose occupations have tipped from being male-dominated to female-dominated, thus re-segregating quite rapidly (Reskin and Roos 1990). In addition, jobs can vary in their gender assignment from society to society (Jacobsen 2006). Thus the maintenance of gender segregation in and of itself appears to be even more fundamental than essentialist arguments regarding differential ability and/or job preferences can explain.

A notable pattern is that female-dominated jobs tend to pay less, even to the men in them, than do 'comparable' male-dominated jobs. Thus occupational segregation is linked with lower pay for women. This relationship could arise through various mechanisms. If women are crowded into a smaller set of occupations through hiring discrimination, crowding will lead to lower pay for women if labour demand does not adjust across occupations. If women are willing to trade off pay for working conditions, crowding into the more desirable occupations means lower pay by choice. However, reducing occupational segregation is neither necessary nor sufficient to raise pay for women; some countries (such as Sweden and

Australia) with higher relative earnings for women also exhibit greater occupational segregation than countries with lower relative earnings for women (such as Japan and the United States) (Jacobsen 2006).

Policies Affecting the Gender Division of Labour

In post-industrial societies, many public and business policies affect the gender division of labour and occupational segregation. Any policy affecting the net wage rate, including taxes on earnings, the deducibility of childcare expenses, or means-tested government benefits, affects the market work-non-market work–leisure trade-off. In general, the asymmetry between taxable income and non-taxable household production produces a bias towards household production. The net effect on behaviour of the large number of relevant policies is unclear.

Few policies have directly aimed at reducing occupational segregation. Affirmative action has had much more notable effects on racial employment than on gender employment patterns. Educational access policies, such as the opening of college and postgraduate programmes to women, have been more important, particularly for increasing women's participation in the professions. However, the focus on access to formal education has not encouraged women to enter those jobs traditionally learned through apprenticeships, such as the crafts and trades; these areas continue to be among the most male-dominated of occupations. Meanwhile, few policies have encouraged men to enter female-dominated occupations such as nursing and childcare, even as shortages of caring labour appear.

Lack of explicit gender-desegregation public policy reflects the continued ambivalence in society regarding the desirability of gender desegregation. This stands in notable contrast to stated beliefs regarding racial desegregation, where separatism has become increasingly spurned. Gender segregation occurs in other social spaces such as sports and schooling. Most amateur sports teams continue to be gender-segregated even as US Title IX legislation and similar actions in other countries increases access to sports for high school and college women. Single-sex schooling persists in a wide range of societies and is even encouraged up through high school, although colleges are mainly co-educational except in the most gender-segregated societies such as Saudi Arabia. This segregation is often couched in terms of improving women's (and sometimes men's) outcomes (that is, through arguing that women perform better in single-sex systems), yet still constitutes an argument for separate spheres. In addition, ambivalence continues towards men raising children, the desirability of outsourced childcare, and thus the desirability of mothers working full-time. Economists and other social scientists have performed a useful service in documenting the extent and nature of gender segregation, but have not yet led a full public debate as to its desirability.

See Also

- ▶ Becker, Gary S. (Born 1930)
- ▶ Family Economics
- ▶ Marriage Markets
- ▶ Social Norms
- ▶ Women's Work and Wages

Bibliography

Akerlof, G., and R. Kranton. 2000. Economics and identity. *Quarterly Journal of Economics* 115: 715–753.

Anker, R. 1998. *Gender and jobs: Sex segregation of occupations in the world*. Geneva: International Labour Office.

Badgett, M., and N. Folbre. 2003. Job gendering: Occupational choice and the marriage market. *Industrial Relations* 42: 270–298.

Baker, M., and J. Jacobsen. 2006. Marriage, specialization, and the gender division of labor. *Journal of Labor Economics* 25: 763–793.

Becker, G. 1971. *The economics of discrimination*, rev. edn. Chicago: University of Chicago.

Becker, G. 1991. *A Treatise on the Family*, enlarged edn. Cambridge, MA/London: Harvard University Press.

Braunstein, E., and N. Folbre. 2001. To honor and obey: Efficiency, inequality, and patriarchal property rights. *Feminist Economics* 7: 25–44.

Brown, C., and M. Corcoran. 1997. Sex-based differences in school content and the male/female wage gap. *Journal of Labor Economics* 15: 431–465.
Corcoran, M., and P. Courant. 1987. Sex-role socialization and occupational segregation: An exploratory investigation. *Journal of Post Keynesian Economics* 9: 330–346.
Duley, M., K. Sinclair, and M. Edwards. 1986. Biology versus culture. In *The cross-cultural study of women: A comprehensive guide*, ed. M. Duley and M. Edwards. New York: Feminist Press, City University of New York.
Elul, R., J. Silva-Reus, and O. Volij. 2002. Will you marry me? A perspective on the gender gap. *Journal of Economic Behavior and Organization* 49: 549–572.
Engels, F. 1884. *The origin of the family, private property, and the state*. Hottingen-Zurich: Swiss Co-operative Printing Association.
Engineer, M., and L. Welling. 1999. Human capital, true love, and gender roles: Is sex destiny? *Journal of Economic Behavior and Organization* 40: 155–178.
Fausto-Sterling, A. 1985. *Myths of gender: Biological theories about women and men*. New York: Basic Books.
Fawcett, M. 1892. Mr. Sidney Webb's article on women's wages. *Economic Journal* 2: 173–176.
Folbre, N. 1995. 'Holding hands at midnight': The paradox of caring labor. *Feminist Economics* 1: 73–92.
Galor, O., and D. Weil. 1996. The gender gap, fertility, and growth. *American Economic Review* 86: 374–387.
Greenwood, J., A. Seshadri, and M. Yorukoglu. 2005. Engines of liberation. *Review of Economic Studies* 72: 109–133.
Gronau, R. 1988. Sex-related wage differentials and women's interrupted labor careers – The chicken or the egg. *Journal of Labor Economics* 6: 277–301.
Grossbard-Shechtman, S. 1993. *On the economics of marriage: A theory of marriage, labor, and divorce*. Boulder/Oxford: Westview Press.
Hadfield, G. 1999. A coordination model of the sexual division of labor. *Journal of Economic Behavior and Organization* 40: 125–153.
Hartmann, H. 1976. Capitalism, patriarchy, and the division of labor. *Signs* 1((3), pt. 2): 137–169.
Humphries, J. 1991. The sexual division of labor and social control: An interpretation. *Review of Radical Political Economics* 23: 269–296.
Jacobsen, J. 2006. *The economics of gender*. 3rd ed. Malden: Blackwell.
Lundberg, S., and R. Pollak. 1994. Noncooperative bargaining models of marriage. *American Economic Review* 84: 132–137.
Lundberg, S., and R. Startz. 1983. Private discrimination and social intervention in competitive labor markets. *American Economic Review* 73: 340–347.
Matthaei, J. 1995. The sexual division of labor, sexuality, and lesbian/gay liberation: Towards a marxist-feminist analysis of sexuality in U.S. capitalism. *Review of Radical Political Economics* 27: 1–37.
Reskin, B., and P. Roos. 1990. *Job queues, gender queues: Explaining women's inroads into male occupations*. Philadelphia: Temple University Press.
Siow, A. 1998. Differential fecundity, markets, and gender roles. *Journal of Political Economy* 106: 334–354.

General Equilibrium

Lionel W. McKenzie

Abstract

Unlike partial equilibrium theory, general equilibrium theory treats as constant only non-economic influences and embraces all sales and purchases of all agents involved in exchanges. It implies that all subsets of agents are in equilibrium and that all individual agents are in equilibrium. The development of a formal general equilibrium theory in mathematical terms was initiated in the 19th century by Walras, who moved from a model of an exchange economy to an equilibrium with production. It was completed in the 1950s by McKenzie, who formalized Walrasian theory, and by Arrow and Debreu, who formalized Hicksian theory.

Keywords

Arrow, K.; Arrow–Debreu model; Competitive equilibrium; Convexity; Cournot, A.; Debreu, G.; Equilibrium of production; Equilibrium over time; Existence of general equilibrium; General equilibrium; Hicks, J.; Jevons, W.; Marshall, A.; Menger, C.; Mill, J. S.; Monetary equilibrium; Overlapping generations model of general equilibrium; Pareto efficiency; Pareto, V.; Partial equilibrium; Poinsot, L.; Rational expectations; Samuelson, P.; Smith, A.; Temporary equilibrium; von Neumann's Law; Wald's Law; Walras, L.; Walras's Law; Walras's theory of investment

JEL Classifications

D5

General equilibrium theory is in contrast with partial equilibrium theory where some specified part of an economy is analysed while the influences impinging on this sector from the rest of the economy are held constant. In general equilibrium the influences which are treated as constant are those which are considered to be noneconomic and thus beyond the range of economic analysis. Of course, this does not guarantee that these influences will in fact remain constant when the economic factors change, and the usefulness of economic analysis for predictive purposes may depend on to what degree influences treated as noneconomic are really independent of the economic variables.

The institution whose phenomena are the primary subject matter of economic analysis is the market, made up of a group of economic agents who buy and sell goods and services to one another. In partial equilibrium theory the group of agents may be confined to those who are involved in one industry, either buying or selling its product or buying or selling the materials and productive services used in making its product. However, in general equilibrium theory all the agents involved in exchanges with each other should ideally be included and all their sales and purchases should be allowed for. However, it may happen that the activities of many agents are only treated in the aggregate and the list of goods and services may be reduced by aggregation. The aggregation of agents and commodities into a few categories is especially important when general equilibrium theory is applied to special areas of public policy such as the government budget, money and banking, or foreign trade. Much of the theory developed for these subjects is general equilibrium theory in aggregated form.

The general equilibrium implies that all subsets of agents are in equilibrium and in particular that all individual agents are in equilibrium. The conscious development of a formal general equilibrium theory stated in mathematical terms seems to have been inspired by a formal theory of the equilibrium of the individual consumer faced with a given set of trading opportunities or prices. This theory was developed by the marginal utility, or neo-classical, school of economists in the third quarter of the nineteenth century, independently, by Gossen (1854), Jevons (1871), and Walras (1874–7), who used mathematical notations, and by Menger (1871) who did not. The step was taken in the most effective way by Walras.

The Equilibrium of an Exchange Economy

Walras assumed that the utility derived from the consumption of a good was given as a function of the amount of that good alone that was consumed and independent of the amounts consumed of other goods. He also assumed that the first derivative of the utility function was positive and decreasing up to a point of satiation when one exists. He then gave a rigorous derivation of the demand for a good by a consumer from the maximization of utility subject to a budget constraint. The demand functions give the equilibrium quantities traded by the consumer as a function of market prices. As Walras saw, this is a crucial step in the development of a general equilibrium theory for an economy. It has remained in a generalized form the cornerstone of general equilibrium theory since Walras.

The simplest problem of general equilibrium arises in the theory of the exchange economy without production. In this economy the budget constraint of the trader is established by his initial stocks and the list of prices. Then the individual demand function represents the equilibrium of the single trader in face of a given price system. The market demand function is the sum of the individual demand functions, and the equilibrium of the market occurs at a price for which the sum of demands, including offers as negative demands, is equal to 0 for each good, or, if free disposal is allowed, is not positive for any good. This idea was expressed in classical economic theory by the equality of supply and demand in each market, but its expression in a set of equations to be satisfied by the list of equilibrium prices was due to Walras, although Cournot (1838) had foreshadowed the Walrasian analysis in his discussion of the

international flow of money and Mill (1848) in his discussion of foreign trade.

Suppose there are n goods to be traded and there are m traders. Let w_i^h be the quantity of the ith good held initially by the hth trader. Let $u^h(x)$ where $x = (x_1,...,x_n)$ be the utility to the hth trader of possessing the quantities $x_1,...,x_n$, of the n goods traded. Then the hth trader is in equilibrium at the prices $p = (p_1,..., p_n)$ and the quantities x^h if $u^h(x)$ is a maximum at x^h over all values of x which satisfy $\sum_1^n p_i x_i \leq \sum_1^n p_i w_i^h$. If smoothness and concavity conditions are met by the utility function, and the goods are divisible, the maximizing x will be unique and will define a function $f^h(p)$ over an appropriate price domain. Since the set of commodity bundles x at which the utility function is maximized does not change when the prices p are multiplied by a positive scalar, this function will satisfy $f^h(p) = f^h(\alpha p)$ for $\alpha > 0$.

The market demand function is $f(p) = \sum_1^m f^h(p)$. Then the market equilibrium for a trading economy is given by a price vector p and an allocation of goods $(x^1, ...,x^m)$ such that $x^h = f^h(p)$ and $\sum_1^m x^h = \sum_1^m w^h$, or, assuming free disposal, $\sum_1^m x^h \leq \sum_1^m w^h$. The first condition expresses the equilibrium of the individual trader and the second condition is the equality of supply and demand. Thus there are n scalar equations $\sum_{h=1}^m f_i^h\left(p = \sum_{h=1}^m w_i^h\right)$ to determine the n equilibrium prices p_i. The given data are the consumer tastes, expressed in the utility functions u^h, and the initial stocks of goods w^h.

It is clear that the market demand function satisfies the homogeneity condition $f(\alpha p) = f(p)$ for $\alpha > 0$. Thus equilibrium prices are only determined up to multiplication by a positive number. This reflects the fact that the equilibrium of the consumer is not affected if prices are multiplied by α and market equilibrium is the simultaneous equilibrium of all consumers at the same prices. It is often convenient to adopt some normalization of prices. Walras chooses a good whose price is known to be positive in equilibrium and gives this good, which he calls the numeraire, the price 1. Another convention which is useful when free disposal is assumed, so that prices are necessarily non-negative, is to choose p such that $\sum_1^n p_i = 1$. Then the domain of definition for the demand functions may be taken to be all p such that $p_i \geq 0$ and $\sum_1^n p_i = 1$.

There is an analogy between the equilibrium of the trading economy and the equilibrium of mechanical forces. Indeed, one of the inspirations for the theory of Walras appears to have been a treatise on statics by Poinsot (1803, 1842). According to the principle of virtual work an infinitesimal displacement of a mechanical system, which is at equilibrium under the stress of forces and subject to constraints, does no work. In the economy at equilibrium an infinitesimal displacement of the allocation of goods $(x^1,..., x^m)$ cannot increase the utility of one trader unless it reduces the utility of another. This is an easy implication of the fact that utility is maximized over the budget constraint, provided no one is saturated. This means that a new allocation to a trader cannot preserve his utility level if its value at the equilibrium prices falls. On the other hand, the utility level of a trader cannot increase unless his allocation becomes more valuable at the equilibrium prices. But then the new allocations $x^{h'}$ would satisfy

$$\sum_{h=1}^{m}\sum_{i=1}^{n} p_i x_i^{h'} > \sum_{h=1}^{m}\sum_{i=1}^{n} p_i w_i^h$$

which is impossible since the total allocation cannot exceed the total supply of goods. Indeed, if each trader holds all goods in his equilibrium allocation and the utility functions are differentiable, which implies that goods are divisible, an infinitesimal reallocation would have no effect on utility levels if it has no effect on the levels of individual budgets. This property of market equilibrium was first recognized by Pareto (1909), and an allocation of goods with the property that no displacement of it can benefit one consumer unless it harms another is said to be Pareto optimal. The implication from competitive equilibrium to Pareto optimality requires that no consumer be locally satiated. It is also true that a Pareto optimal allocation may be realized as a competitive equilibrium given an appropriate distribution of initial stocks but the conditions are

more severe. The first general theorems were proved by Arrow (1951).

Equilibrium with Production

The next step in developing the general equilibrium of an economy is to introduce production under the condition that the output matures without a lapse of time. This step was taken by Walras, who introduced linear activities which list the quantities of productive services required to produce one unit of a good. There may be many alternative activities for the production of any given good and a choice is made among them in order to minimize the cost of production at given market prices. Let $z = (z_1, \ldots, z_r)$ be a list of quantities of productive services and let $g^i(z)$, $i = 1,\ldots, n$, be production functions for the n goods. Since linear activities are assumed, the production functions will satisfy $\alpha g^i(z) = g^i(\alpha z)$. In particular, we may consider the unit isoquant or the set A_i such that $g^i(z) = 1$ for z in A_i. Then the activities which minimize cost at given prices q are represented by production coefficients $a^i(q)$, contained in A_i, where $q'a^i(q) \leq q'z$ for z in A_i. Equilibrium in the production sector is given by price vectors p and q and activity vectors $a^i(q)$ where $p_i \leq \sum_{j=1}^{r} q_j a_j^i(q)$ for all i and equality holds if the ith good is produced.

In an equilibrium of the production sector any quantities y of outputs may be produced provided quantities z of productive services are available where $z_j = \sum_{i=1}^{n} y_i a_j^i(q)$. In order to include the productive sector in a market equilibrium the utility functions of consumers must be extended to include productive services among their arguments. They may be written $u^h(x, z)$. If were interpret x_i as the quantity of a good traded rather than the quantity consumed, the initial stocks may be suppressed. This is convenient since it is not clear how initial stocks of labour services can be specified. Then the individual consumer is in equilibrium given prices p and q for goods and productive services when the quantities traded (x^h, z^h) maximize $u^h(x,z)$over all (x, z) such that $\sum_1^n p_i x_i - \sum_1^r q_i z_i \leq 0$. The maximizing quantities need not be unique in general, so it is necessary to represent demand by a correspondence that takes a set of trades as its value and write $(x^h,z^h) \in f^h(p, q)$ when (x^h, z^h) is a maximizer given prices p and q.

As before, market equilibrium is achieved when all economic agents are in equilibrium at the same prices and supply is equal to demand. Since risk is not present in this economy, the productive services involved in organizing production need not be given a distinguished role. Activities may be treated as conducted by the whole set of owners of the productive services involved in them. Then if it should happen that $p_i > \sum_{j=1}^{r} q_j a_j^i(q)$ for the ith good, there will be an opportunity for some owners of productive services to earn larger returns producing the ith good than those prevailing generally as given by q. Thus productive services will leave other activities and flow to this activity, so equilibrium does not obtain for owners of productive services. This equilibrium now requires, on the one hand, equilibrium of each economic agent as consumer of goods and provider of productive services, that is, $(x^h, z^h) \in f^h(p, q)$, and, on the other hand, equilibrium of each economic agent as a participant in production, that is, $p_i \leq \sum_{j=1}^{r} q_j a_j^i(q)$, with equality if the ith good is produced. However, market equilibrium also requires that $\sum_{h=1}^{m} z_j^h = \sum_{h=1}^{m} \sum_{i=1}^{r} x^h a_j^i(q)$, that is, the supply of productive services must equal the quantities needed to produce the quantities of goods demanded. As before, if surplus productive services may be freely disposed of, the equality in the last equation may be replaced by an inequality.

The demand functions $f_i^n(p, q)$ and the supply functions $f^h{}_{n+j}(p, q)$ express the equilibrium of the household sector. Therefore, the relation $\sum_{i=1}^{n} p_i f_i^h(p, q) = \sum_{j=1}^{r} q_i f_{n+j}^h(p, q)$holds for all values of p and q in the price domain. Let xi be the amount of the ith good produced and let zj be the amount of the jth factor used in production. Then equilibrium in the production sector implies that

$$\sum_{i=1}^{n} p_i x_i = \sum_{j=1}^{r} \sum_{i=1}^{n} q_j x_i a_j^i = \sum_{j=1}^{r} q_j z_j.$$

G

Let $f(p,q) = \sum_{h=1}^{m} f^h(p,q)$. Then household equilibrium implies $\sum_{i=1}^{n} p_i f_i(p,q) = \sum_{j=1}^{n} q_i f_{n+j}(p,q)$. Let excess demand for a good be $e_i(p,q) = f_i(p,q) - x_i$, and excess demand for a productive service be $e_{n+j}(p,q) = z_j - f_{n+j}(p,q)$. Then equilibrium in the production and household sectors together implies that $\sum_{i=1}^{n} p_i e_i(p,q) + \sum_{j=1}^{n} q_j e_{n+j}(p,q) = 0$, or the value of excess demand is zero whatever price system is set. This relation is referred to as Walras's Law.

If there is free disposal, prices must be non-negative. Otherwise, disposal would be profitable. Also with free disposal the condition for equilibrium of the market is $e(p,q) \leq 0$. Then Walras's Law immediately implies $p_i e_i(p,q) = q_j e_{n+j}(p,q) = 0$, and if any good or productive service is in excess supply in equilibrium, its equilibrium price must be 0. This might be termed Wald's Law, since he made crucial use of it in the first rigorous proof that equilibrium exists in a competitive economy (Wald 1935, 1936).

A production sector composed of activities with single outputs is the model used by Walras, who was responsible for the first fully developed general equilibrium theory. The natural generalization of this model is to introduce more than one output. Then the kth activity is represented by an output vector $b^k = \left(b_1^k, \ldots, b_n^k\right)$ and an input vector for productive services $a^k = \left(a_1^k, \ldots, a_r^k\right)$. Assume that activities may be replicated and are independent of each other. Then if (a^k, b^k) is a possible input-output combination for the kth activity, so is $(\alpha a^k, \alpha b^k)$ where α is any non-negative integer. Indeed, if all inputs and outputs are divisible it is possible for α to take as its value any real number.

This model of the production sector which embraces the transformation of productive services into goods and services is due to Walras in the context of a theory of general equilibrium. It is convenient to think of the market as held periodically to arrange for the delivery of goods and services over a certain basic period of time. This view of the market, which is also a device of Walras, leads to a theory of temporary equilibrium. The theory was further elaborated by Hicks (1939) and in recent years by other authors. In order to explain the demand and supply of products and productive services in the periodic market it is necessary to introduce some assumptions on the formation of expectations for the prices which will prevail in future markets. The simplest assumption is that the prices arrived at in one market are expected to prevail in future markets. This type of expectation formation is sometimes referred to as static expectations. Walras usually appears to assume static expectations. Hicks introduced a notion of elasticity of expectations to allow expectations of future prices to depend on the change of prices from one temporary equilibrium to another. In recent work analysis has proceeded upon more general assumptions, using various formal properties of dependencies between past prices and expected prices. A quite different approach to expectations which enjoys much current popularity is to assume that expectations are correct, at least in a stochastic sense. The rationale of this approach is that any persistent bias in forecasts of future prices implies that there are unexploited opportunities for profit from further trading which eventually should be recognized.

The model of the production sector as a set of potential linear activities was subsequently used by Cassel (1918) in a simplified Walrasian model which preserved the demand functions and the production coefficients but which did not deduce the demand functions from utility functions or preferences. The model was generalized to allow joint production in a special context by von Neumann (1937). It was given a thorough elaboration and analysis in a model where intermediate products are introduced explicitly by Koopmans (1951). In the Walrasian picture intermediate products were eliminated through the combination of activities so that activities were described as transforming productive services directly into final products whether consumer goods and services or capital goods. However, such a description of the economy depends for its relevance on prices which do not change from one temporary equilibrium to another, so that the choice of activities is not changing.

In the general linear model of production it is no longer adequate to treat the choice of activities

as a process of cost minimization given the price vectors p and q. Cost minimization must be replaced by the condition that no activity may offer a profit and no activity which is used in competitive equilibrium may suffer a loss.

This is exactly the condition '*ni benefice ni perte*' which Walras used to define equilibrium in production, initially in a model with fixed coefficients of production. However, this condition was first used in a general production model by von Neumann, so it might be termed von Neumann's Law for an activities model of production. Koopmans explored the relation between efficient production and von Neumann's Law. He established an equivalence between the proposition that an output is efficient and the proposition that prices exist such that von Neumann's Law is satisfied when the activities used are those needed to produce this output.

Moreover, if each good or service is either desired in unlimited quantities or freely disposable the prices must be non-negative. Thus under these demand and supply conditions any competitive equilibrium must include an efficient output from the production sector. The activities approach to the production sector of a competitive economy was used by Wald and then by McKenzie (1954) in proofs of existence for competitive equilibrium. It was also used by Scarf (1973) in an algorithm for finding a competitive equilibrium given the technology, the resources, and the demand functions.

An alternative model of the production sector emphasizes the productive organization or firm rather than the activities or technology. A set of actual or potential firms is given and each firm is endowed with its own set of possible input- output combinations. The set of possible input-output combinations achievable by the economy, independently of resource availabilities, is the sum of the sets of input- output combinations achievable by the firms. The condition for equilibrium in the production sector is that each firm maximizes its profits, that is, the value of the input-output combination over its production possibility set, given the prices of inputs and outputs. This view of production was explicit in a partial equilibrium context in Cournot. It was at least implicit in the work of Marshall (1890) and Pareto, and became quite explicit in a general equilibrium context in the work of Hicks (1939) and Arrow and Debreu (1954).

In the Hicksian model a firm is associated with each economic agent who is a consumer and who may be a worker and owner of resources, but who also may be an entrepreneur. As an entrepreneur he owns a possible production set based on his personal characteristics and perhaps some other non-marketed resources. Of course, most of these individual enterprises will be inactive. A difficulty with this model is that it seems unrealistic to treat the entrepreneur as a profit maximizer unless all the resources which he himself supplies have market prices so that they could equally well be bought by him from the market or sold by him to the market. But if that is the case we are back to the concept of the entrepreneur used by Walras and it seems more realistic to refer to activities, which are impersonal, rather than to individual enterprises.

In the model of Arrow and Debreu, which is the first complete general equilibrium model in which the existence of equilibrium was rigorously proved, the production sector is made up of firms which are described as joint stock companies. Each firm has a production possibility set based on resources which it owns and the ownership of the firm is spread in a prescribed way over a set of consumers. The production sector is in equilibrium when each firm has chosen an input-output combination from its production possibility set which maximizes profit at the market prices. Since the outputs of one firm may be inputs of another and the resort to integrated activities which convert productive services directly into products is not available in a model based on firms, it is convenient to distinguish inputs from outputs by signs rather than by lists. Let Y_j denote the production possibility set of the jth firm, and let $y = (y_1,..., y_n)$ denote an element of this set. There are n goods and services in the economy, and $y_i < 0$ denotes an input, while $y_1 > 0$ denotes an output. Let y^j be the input-output vector of the jth firm. Then equilibrium in the production sector requires that the condition $p\,y^j \geq p\,y$ for all $y \in Y_j$ holds for all j, where j indexes the set of firms, and p is the market price vector.

The Arrow–Debreu approach to the production sector involves a major difficulty. It is not well adapted to handle the formation of new firms and the dissolution of old ones. If firms are based on the assembly of a set of resources jointly owned by the shareholders, it becomes critical to give the principle which underlies such an assembly. If the firm's resources are priced and traded, so the firm's production may be treated like an activity, there is no difficulty since von Neumann's Law may be applied. Otherwise, the rules governing the entry and exit of firms are unclear. The problem is similar to the general problem of coalition formation in the theory of cooperative games.

A Formal Model

A formal model of the competitive economy, presented in the form of a series of axioms, was developed in the 1950s. It was intended that the axioms should be interpretable to apply to real economic systems, albeit in some approximate sense. However, as a formal mathematical model the implications of the axioms could be developed independently of the applications. The selection of axioms was influenced by the possibility of making useful interpretations, but also by the facility with which results can be derived.

Two closely related sets of assumptions were developed. One, developed primarily by McKenzie (1959), is a formalization of the Walrasian theory and uses a linear model of production. The other, developed primarily by Arrow and Debreu, is a formalization of the Hicksian theory where the production sector is described as an assembly of firms. On the side of consumers and the market there are no significant differences at a fundamental level, although there are sometimes differences of approach. A history of the problem of existence of equilibrium for the formal models may be found in Weintraub (1983).

In the fully developed McKenzie model (see McKenzie 1981) two assumptions are made for the consumption sector, two for the production sector, and two assumptions relate the consumption and production sectors. On the consumption side there is a finite number m of consumers indexed by h, and each consumer has a set X_h of trades which are feasible for him. There are n goods and the sets X_h are contained in R^n, the n dimensional Euclidean space. The convention is used that quantities supplied by consumers are negative and quantities received by consumers are positive. The consumer has preferences defined on X_h by a correspondence P_h. The preference correspondence P_h takes as its value at $x \in X_h$ the subset of X_h each of whose members is preferred to x. This subset may be empty. The assumptions on the consumers which hold for all h, are:

(1) X_h is convex, closed and bounded below.
(2) P_h is open valued relative to X_h and lower semi-continuous. Also x is not in convex hull $P_h(x)$.

Convexity of X_h implies that a good is divisible if someone can consume it in more than one quantity. X_h bounded below means that the consumer is not able to supply an indefinite quantity of any good.

Closedness and boundedness are needed to provide compact feasible sets.

On the production side there is an activities model with no limitation on the number of activities. The activities are linear and give rise to a possible production set Y contained in R^n. If $y \in Y$, the negative components of y denote quantities of inputs and the positive components denote quantities of outputs.

The assumptions on Y are:

(3) Y is a closed convex cone.
(4) $Y \cap R^n_+ = \{0\}$. R^n_+ is the set of non-negative vectors in R^n.

That Y is a convex cone is equivalent to the production set being generated by linear activities. It means that if y and y' are producible, that is, elements of Y, then $\alpha y + \beta y'$ is also producible, that is, anelement of Y, for any non- negative numbers α and β. Thus producible goods are divisible. Closedness is needed for the compactness of the feasible set. Assumption (4) is not restrictive. It is a recognition that goods which are never scarce are irrelevant to problems of economizing.

Finally two assumptions relate the consumption sector and the production sector. Let X be the total possible consumption set, that is, $X = \sum_{h=1}^{m} X_h$. The first relation is

(5) Relative interior $X \cap$ relative interior $Y \neq \emptyset$.

Here the relative interior of a set is relative to the smallest linear subspace that contains it. This assumption insures that someone has income at any price vector which is consistent with equilibrium in the production sector, that is, satisfies von Neumann's Law. The second relation is an assumption that the economy is irreducible. Let I_1 and I_2 refer to nonempty subsets of consumers such that $I_1 \cup I_2$ includes all consumers and $I_1 \cap I_2 = \emptyset$. Let $X^1 = \sum X_h$ for $h \in I_1$, and similarly for I_2. Let $\overline{X}_h$ be the convex hull of X_h, and the origin of Rn. The irreducibility assumption is

(6) However $I_1$and I_2 may be selected, if $x^1 = y - x^2$ with $x^1 \in X^1$, $y \in Y$and $x^2 \in \overline{X}^2$, then there is also $\tilde{y} \in Y$ and $w \in \overline{X}^2$, such that $\tilde{x}^1 = \tilde{y} - x^2 - w$ and $\tilde{x}^1 \in P\left(x^h\right)$ for all $h \in I_1$.

Assumption (6) guarantees that everyone has income if anyone has income. The meaning of having income is that the consumer is able to reduce his spending at the market price vector below the cost of his allocation and remain within his possible consumption set X_h.

Competitive equilibrium is defined by a price vector p, an output vector y, and vectors $x^1,...,x^m$ of consumer trades. There is equilibrium in the production sector if von Neumann's Law holds, that is:

(I) $y \in Y$ and $p \cdot y = 0$, and for any $y' \in Y, p \cdot y' \leq 0$.

When y satisfies (I) it is not possible for the owners of inputs to withdraw them from activities where they are being used and employ them in other activities, whether in use or not, so that the receipts from the resulting outputs allow some inputs to earn larger returns while none of them earns less. This is the same condition for equilibrium in production that was given by Walras, or, for that matter, by Adam Smith (1776).

There is equilibrium in the consumer sector if the x^h satisfy

(II) $x^h \in X_h$ and $p \cdot x^h \leq 0$, and $p \cdot z > 0$ for any $z \in P_h(x^h)$, $h = 1, \ldots, m$. When x^h satisfies condition (II), there is no preferred bundle of goods, including goods or services that are supplied by the consumer, which is available to him under his budget constraint. This is essentially the same condition used by Walras, except that he assumed that preferences could be represented by a strictly concave utility function. Thus he is able to refer to maximization of the utility function over the budget set uniquely at x^h.

Finally, there is market equilibrium when

(III) $\sum_{h=1}^{m} x^h = y$.

This is the condition that markets clear which was used by Walras.

If there is free disposal, Wald's Law may be derived directly from equilibrium in the production sector. The possibility of free disposal is recognized by the inclusion of disposal activities in the production cone, that is, an activity y^i for $i = 1,...,n$ which has $y_i^i = -1$ and $y_j^i = 0$ for $j \neq i$. The condition $p \cdot y^i \leq 0$ implies that $p_i \geq 0$ must hold. Then if disposal occurs the condition $p \cdot y^i = 0$ implies that $p_i = 0$.

On the basis of Assumptions 1 through 6 it is possible to prove that a competitive equilibrium exists. This was first achieved in a model with assumptions for the demand sector put directly on preferences, in the manner of Walras, by Arrow and Debreu. At the same time McKenzie proved existence for a model with assumptions put on the demand functions rather than directly on preferences. Also McKenzie assumed a linear technology rather than a set of firms. This was a generalization of a model of Wald in which joint production was absent and the very special assumption was made that the market demand functions satisfied the weak axiom of revealed preference. The weak axiom says if x is demanded at p and x' at p', then $p \cdot x' \leq p \cdot x$ implies that $p' \cdot x < p' \cdot x$. This is a consistency requirement on choice under budget constraints. Wald's assumption was a deep insight. He anticipated the

G

statement of this principle by Samuelson (e.g. 1947) who applied it to the demand of the individual consumer to derive most of the propositions of demand theory. Wald showed that the weak axiom assumed for the market leads to uniqueness of equilibrium. Subsequently it was shown by Arrow and Hurwicz (1958) that the weak axiom is implied by the assumption that all goods are gross substitutes. They also proved that the weak axiom confined to a comparison of choices between the equilibrium prices and other prices implies the global stability of a process of price adjustment in which the prices of goods are increased if excess demand exists and lowered if excess supply exists. Wald (1936) wrote another paper on equilibrium in an exchange market which used assumptions closer to those of Arrow and Debreu, but this paper unfortunately was lost.

The only important distinction between the approach of Arrow and Debreu (see Debreu 1962) and the approach expressed in Assumptions 1 through 6 is the use of a set of firms rather than a set of activities to generate the production set.

Mathematically, through the introduction of entrepreneurial factors the approaches can be reconciled. However, the intentions of the two approaches are quite different. The linear model is intended to represent free entry into any line of production by cooperating factors, however organized in a legal sense, where economies of scale are sufficiently small to allow approximate linearity to be achieved by the multiplication of producing units. The lumpiness which is present is compared to that resulting from goods which are in fact indivisible, although they are treated as divisible. This leads to a reasonable approximation to real markets only if units are small compared with the levels of trade. This view of the competitive economy is consistent with the analysis of Marshall as well as Walras. Of course, it has to be recognized that in real economies some sectors cannot be approximated in this way. However, when linearity becomes a bad approximation to the production sector, convexity has in all likelihood become an equally bad approximation to the production sets of firms.

Recently an explicit modelling of the approach of the firms economy to the activities economy has been given by Novshek and Sonnenschein (1980). They use the model of quantity adjusting firms developed in a partial equilibrium context by Cournot to find an equilibrium for the firms economy. Then they let the firm size shrink and show that the Walrasian equilibrium of an activities economy is approached in the limit.

Two Interpretations of the Formal Model

Two basic interpretations of the general equilibrium model were described by Hicks and referred to as the spot economy and the futures economy. The spot economy is a market held on 'Monday' at which all transactions are arranged that involve delivery during the 'week'. This is the economy described by Walras. The equilibrium of the spot economy is called temporary equilibrium in the modern literature. Some effort has been devoted to an analysis of the path followed by such an economy through a succession of temporary equilibria. The role of expectations in the spot economy is critical, as Hicks recognized.

The futures economy on the other hand has a single market in which all future transactions are negotiated at once. Hicks does not treat this economy in detail, but turns to a sequence of spot markets with trading that is guided by expectations. In the futures economy goods available in different periods would be treated as different goods, so that the number of goods would be finite only if the economic horizon is finite. If there is perfect foresight the futures economy is a reasonable alternative and there is not reason why markets should reopen. However, when the future is uncertain and the available futures contracts are for sure delivery, or at least do not exist in sufficient variety to take account of all contingencies, there is no assurance that the contracts entered into will remain desirable or indeed can be executed. For this reason Hicks chose to do a dynamic analysis of a sequence of temporary equilibria in the main body of his work.

In order to avoid the problem of the feasibility of plans and the need to reopen markets, Debreu

(1959) following a lead of Arrow (1953) introduced a specification of goods by the event in which they are made available. The set of events would have to discriminate all the circumstances that might make delivery impossible or undesirable, so there would be no motive for traders to reopen markets. Despite this complexity, it is a consistent model which may have relevance to the real world. In order to keep the set of goods finite they assume a finite horizon and a finite set of events, in addition to assuming a finite list of goods in terms of location and physical characteristics.

With this interpretation of the formal model there is no room for borrowing and lending since payments are cleared only once, at the beginning of time. Uncertainty is present since there is no assumption that the event realized at any future time is known. Rather it will be revealed when the time arrives. There is no reason for spot markets to arise since the transactions which have been made for the future event that is revealed are the ones each trader desired at the prices paid in those circumstances. Thus if a spot market were opened no transactions would take place.

Of course it is idealization to suppose that all relevant events could be described in advance, or, if they could, that it would be feasible to establish markets discriminating between them. An alternative is to use a succession of markets in which temporary equilibria are established while some trading in futures contracts takes place. However, the limiting cases of the pure spot economy or the pure futures economy have an analytical tractability that the mixed cases lack and for this reason they remain of great importance.

Temporary Equilibrium

Once a sequence of markets is contemplated, rather than a single comprehensive market, plans for future trades become relevant and, therefore, expectations of the prices at which they can be made. Also money stocks and loans become useful in making financial preparations for the trading that is planned. Also, if there may be forward trading as well as spot trading, arbitrage is possible, and speculative trading arises which expresses disagreement among consumers about probable price levels on future spot and forward markets.

These complications were handled by Walras without an explicit analysis of demand by consumers for goods in the future using utility functions in which these goods appear. Rather he reduces the demand for future goods to a demand for assets in general which would provide the means for future purchases. On the other hand, he carefully distinguishes between stocks of goods and their services, and the investments of the consumer are treated as if they were made directly in the stocks of goods whose services are sold to the entrepreneurs, or directly to consumers in the case of services of consumer goods.

The spirit of this analysis is to choose a period short enough that it is not too great a distortion of reality to suppose that all trades for this period can be concluded in advance as in the Arrow–Debreu model for the entire horizon, but the forms of industrial organization are abstracted from, so that attention may be concentrated on the productive activities and the ultimate beneficial owners of the resources whose services are used in them. Also to give the future some role in the decisions of the consumers but not a role requiring detailed analysis, Walras assumed that present market prices are expected to persist. In contrast, Hicks and Arrow–Debreu deal explicitly with intertemporal planning by firms and consumers. In a succession of markets this allows Hicks to analyse the effects of changes in expectations on the present market prices and the plans of agents.

The theory of Walras provides the most complete and detailed model of temporary general equilibrium that has ever been given, an impressive performance since it was also the first formal model of general equilibrium. He was able to deal with money, production, lending, and capital accumulation, and in his model an interest rate, price levels, and prices of capital goods and their services are all determined. He showed that the system was not overdetermined, and probably not underdetermined either, in that the number of independent functional relationships and the number of economic quantities to be determined are

equal. He was not able to give a proof that an equilibrium in non-negative real variables exists for his model. However, proofs have since been given for simplified versions of it.

A fundamental difference between temporary equilibrium and equilibrium over a horizon is that part of the consumers' demand for goods in the temporary equilibrium is intended for investment rather than for consumption within the period while in the economy of the classical existence theorem consumers' demand is entirely aimed at consumption within the horizon. This raises two problems. One is to distinguish between resources devoted to this period's consumption and resources reserved for the support of consumption in future periods. The other is to explain how the decision to reserve a certain quantity of resources for future use is made.

Walras went further to make the distinction between current and future use than any of his successors. They, on the other hand, have done much more analysis of the relation between investment and expectations. The Walrasian assumption on expectations was usually to project the prices arrived at in the current market into the future. This assumption is only appropriate for a stationary, or a steadily progressive, state of the economy. Of course, it has often been remarked that it is only in these conditions that expectations are likely to be correct.

Walras distinguished between consumption goods and services which are consumed in one use and consumption goods which are in effect capital goods providing consumer services, that is, having more than one use. Among the consumption goods which serve as capital goods he included consumption goods which are held in stocks to provide, as Walras put it, services of availability. Thus part of a person's income for a period may be invested in new stocks of consumer goods as well as in capital goods which are intended for use in productive activities. By the same token some of the productive activities which occur may occur in the household rather than in the factory, and these should satisfy the same profit conditions as the productive activities that occur in the firms.

The Walrasian approach to temporary equilibrium is entirely appropriate only to steady states where underlying circumstances, technology, tastes, and resources are constant, perhaps with capital stocks and population expanding at uniform rates.

Then the comparative statics that can be done is a comparison of different steady states. On the other hand, in the Hicksian model where expectations of price changes are allowed, it is possible to consider the effect on the temporary equilibrium of changes in price expectations which need not duplicate changes in current prices.

However, the approach of Walras allows him to ignore the consumer's portfolio problem and treat the consumer as only making a saving decision, since all assets of equal value are treated as indifferent with equal rates of return after allowing for depreciation and insurance costs. When there is uncertainty, the treatment of all assets as indifferent in this fashion is not justified even by the mean variance theory of portfolio selection. The variances and covariances of asset returns must be taken into account. Thus Walras's theory of investment requires that expectations be held with certainty, although he only explicitly assumes certainty within the horizon of a single period, after allowing for fully insurable risks.

There are two features of the Walrasian theory of investment which are quite effective, even by modern standards. One is the analysis of the demand for money. Money is needed during the period to make payments which are planned in advance and the cost of this money service is simply the interest on a loan of that amount for the period. This is very close to the treatment of the demand for money for transactions purposes in modern theory. The demand for money as an asset is merged with the general demand for assets, since any net money balance at the end of the period will be expected to be lent at the current interest rate for the next period, either to others or implicitly to oneself. This represents a cash balance approach to monetary theory where cash balances are only wanted for transactions purposes. It leads to a strict quantity theory of the price of money in terms of other goods in comparisons between steady states.

The second effective feature of Walras's theory of investment is the recognition that the cost of investment goods will depend on the level of investment, since in the general equilibrium high levels of investment will raise the prices of the productive services needed to produce investment goods and thus the prices of the investment goods themselves. In this way the Walrasian theory takes account of the distinction between the marginal efficiency of investment and the marginal efficiency of capital familiar in the Keynesian literature, as well as the modern notion of the cost of adjustment resulting from an increase in the level of investment.

The two main deficiencies of the Walrasian theory of temporary equilibrium are its lack of an analysis of the demand for assets in general in terms of the future consumption streams that the assets are expected to support and the expected utility they promise to yield, and its lack of an analysis of the demand for particular assets in terms of the distribution of their expected returns.

The neglect of future plans for consumption in determining current demand was addressed by Hicks. He did not suppose that consumers make detailed plans but that they form vague plans and expectations of future prices, which still allow some comparative statics methods to be applied in estimating the effect on current demand of changes in current or future expected prices.

Since firms are recognized explicitly in Hicks's model, they are also represented as making plans for future inputs and outputs in the light of price expectations, which in his case can be identified with the expectations of individuals who become entrepreneurs. The equilibrium of such a model in one period is a set of prices for all the goods and services traded in the market of that period such that the demand for each good or service, including any contract for future delivery that happens to be traded, equals the supply.

Hicks assumes that each consumer and each firm in its planning applies actual or expected interest rates to discount expected future prices to the present so that the problem of maximizing utility for the consumer, or present value for the firm, does not differ, in principle, from the static problem. However, he must assume that agents are risk neutral or in any case that distributions of prices may be replaced by single prices, or certainty equivalents. Thus he is no more able than Walras to analyse how the value of an asset is influenced by the distribution of its returns. But he is able to consider how changes in current prices influence expected future prices, when expected future prices do not necessarily change by the same amounts. This may be the most significant advance made by Hicks beyond Walras, together with the corollary of planning by firms and consumers for a future that involves expected prices changes.

Expectations in Temporary Equilibrium

A natural way to generalize the Hicksian model and one which has been followed in recent years, for example, by Grandmont, is to impute to each trader an expectation function which gives a probability distribution over future prices, and perhaps over other relevant variables, both market and environmental, as functions of previous values taken by the same variables. Then assuming that each trader has a criterion by which he can choose an optimal trade plan given his expectations, he will determine an excess demand as a function of current prices. Then equilibrium is achieved if there is market clearing at the current prices. Since in the Walrasian or Hicksian model there are two kinds of traders, consumers and entrepreneurs or firms, criteria must be found for each kind of trader.

The criterion for the consumer is rather easily arrived at. It is assumed that each consumer has a von Neumann-Morgenstern utility function, so that any current trade can be evaluated in terms of the expected utility which it makes possible. The utility in turn is derived from the utility of the various possible consumption streams multiplied by their probabilities of occurrence. Of course, these consumption streams and their probabilities logically underlie the expected utilities but they cannot be known to the consumer in detail. The probability distribution on consumption streams is induced by the probability distribution on prices and environmental variables, together with the

current trade of the consumer and his plans for future trades, which are in turn contingent on the prices and environmental variables realized in the future. As Hicks points out the consumer may only try to plan levels of spending and certain large expenditures for the future. Particular price expectations will affect these plans and current spending, in total as well as on specific items. What is needed for the theory is to express consumer's demand finally as a function of current prices so that the condition of market clearing will characterize equilibrium prices. The logic of this analysis is entirely compatible with the methods of Walras, given stationary conditions for tastes, technology, and resources. In simple models it can be spelled out in detail.

On the other hand, there is little agreement on an appropriate criterion for the firm. The difficulty arises that the firm is usually owned by many consumers whose preferences and probability beliefs differ. The consumer does not own capital goods directly but only stock in firms. Moreover, the firms make investment plans and plan their dividend streams in considerable independence of their owners. Walras abstracts from these difficulties in his formal development by two means. First, he treats the consumer as the owner of capital goods which are rented to the entrepreneur. Second, he values the capital goods on the assumption that prices of productive services, interest rates, depreciation rates, and insurance rates will be constant in the future. Given the prices of the productive services arbitrage in the market for capital goods results in a uniform ratio between the net rental of the capital goods, or the prices of their productive services less depreciation and insurance charges, and the prices of the capital goods. In Walras's notation $P_k = p_k/(i + \mu_k + v_k)$ where k indexes capital goods, P_k is the price of the capital good, p_k is the price of its service, i is the interest rate per period, $\mu_k P_k$ is the depreciation change per period, and $v_k P_k$ is the insurance charge per period. In equilibrium the consumer will be indifferent between capital goods in making investments since they all promise equally attractive returns. This also applies in a similar way to investments in circulating capital or in loans.

Hicks adapts the Walrasian viewpoint to a model in which expectations are point valued but not static by imputing to the entrepreneur, who now owns the capital goods, a plan of inputs, including initial stocks, and outputs, including terminal stocks, whose values are discounted back to the present. Then the entrepreneur chooses a plan with the largest discounted value. In this case the firm achieves maximum value in the eyes of its owner. Radner (1972) adapts the Hicksian viewpoint to a model in which point estimates of future prices are not a sufficient basis for decisions. In a temporary equilibrium model his approach imputes to each firm a von Neumann-Morgenstern utility function over alternative dividend streams. This would imply an expected utility for alternative investments in the current period in the same way that the utility of alternative consumption plans implies expected utility for current spending by the consumer.

On the other hand, by use of the stock market it is possible to bring consumers into the decision-making of firms. The firm's criterion is then to choose a plan of production and investment which leads to a maximum value for its shares on the stock market. It can be argued that if the firm chooses a plan which fails to maximize its value in the stock market the stock market will not be in equilibrium, since there is a profitable arbitrage opportunity for someone to buy controlling interest in the firm and revise its planning.

Existence theorems for temporary equilibrium have been proved in many special cases, particularly for trading economies where production does not enter and the number of periods is taken to be finite. Typically the method of proof parallels a method of proof for the model with complete markets, that is, appropriate continuity properties for individual, and thus market, excess demand functions are proved for the goods and services, and the futures contracts, if any, which are traded in the current period. The application of a fixed point theorem completes the proof that a price system exists which results in market clearing, that is, puts each excess demand function equal to zero. However, some special problems do arise.

Consider a market at the start of period 1 when there are two periods and a second market will be held at the start of period 2. There is uncertainty about the endowment of period 2 and about the spot prices of the second market. All goods are perishable. Suppose there is trading in contracts for current delivery and in forward contracts for delivery in the second period. Let x_1^h, x_2^h be the vectors of goods and services delivered to the hth consumer in periods 1 and 2 respectively. Denote by w_1^h and w_2^h the vectors of endowments for the hth consumer in periods 1 and 2 respectively. Let $\psi^h(p_1, q_1)$ be the expectation function of the hth consumer, that is, the value of ψ^h is a probability distribution of $\left(w_2^h, p_2\right)$, where p_1 and p_2 are the vectors of spot prices in periods 1 and 2, while q_1 is the vector of forward prices in period 1 for sure delivery in period 2. There is a finite set of goods and services in each period and a finite number of consumers each of whom holds positive initial stocks in the first period. The possible consumption sets are $X_1^h = R_+^{n_1}$ and $X_2^h = R_+^{n_2}$, the positive orthants of the respective commodity spaces.

The following assumptions are made for the consumer.

(1) There is a concave and monotone utility function u^h of von Neumann- Morgenstern type, that is, preferences over trades in the first period may be determined by taking the sum of the utilities of the resulting consumption vectors weighted by these probabilities of occurrence.
(2) The expectation function $\psi^h(p_1, q_1)$is continuous in an appropriate sense.
(3) For every (p_1,q_1),$\psi^h(p_1,q_1)$ gives probability 1 to the set of $\left(w_2^h, p_2\right)$for which p_2 is positive.
(4) The support of ψ^h is independent of (p_1, q_1). The convex hull of the projection of the support of ψ^h on the second period price space has a non-empty interior Π^h.

With these assumptions a necessary and sufficient condition for the existence of competitive equilibrium is that the intersection Π of the Π^h not be empty. In other words there must be an open set of spot prices in the second period which all traders believe to have a positive probability of occurrence. Then, if the forward prices q_1, lie in Π and p_1 and q_1 are positive, excess demand is well defined. Let D be the set of (p_1, q_1), satisfying these conditions. As (p_1, q_1) converges to the boundary of D, excess demand diverges to ∞. This happens because preferences are monotone and for $q1$ outside Π unlimited arbitrage becomes profitable to some trader. These results were reached by Green (1973). It should be noted that point expectations are not consistent with the assumption that Π is not empty, unless all traders expect the same prices next period. However, Π might not be needed to bound short sales if other considerations limit the commitments that will be accepted in view of the likelihood that they can be fulfilled.

Money in Temporary Equilibrium

There is little difficulty in introducing money into the temporary equilibrium model. It must be recognized that money serves in at least two capacities, to facilitate exchange, and as an asset with its own prospects for losing or gaining value relative to other goods. In addition it may serve as a numeraire, in terms of which prices are stated. In its capacity as an asset in a market with uncertainty, money may contribute to a diversified portfolio. On the other hand, in its capacity to facilitate exchange money balances will affect the cost of making transactions and thus the stream of consumption which is realizable from given resources. Given his context, where risks are assumed to be insurable, Walras is particularly clear in his treatment of money. If some good other than money serves as numeraire, the price of the service of availability of money is written by Walras as p_m, and the price of money itself as P_m. Then as for any asset the ratio of the net rental to the asset price is equal to the interest rate or $p_m/P_m = i$. Thus if money serves as the numeraire, $P_m = 1$ and $p_m = i$. Although his analysis seems somewhat artificial because uninsurable risks are absent, Walras indicates clearly how cash balances may contribute to productive efficiency and to consumer utility.

If attention is concentrated on the asset role of money, so that the transaction role is neglected, it may be shown that the assumption of static may lead to the absence of equilibrium for the current period. Static expectations imply that the relative prices of present and future goods cannot be changed. Therefore, price changes leading to inter-temporal substitution are prevented. Only the wealth effects of price changes have free play since price level decreases raise the value of the money stock and conversely for increases. However, as Grandmont (1983) has demonstrated, these real balance effects may be insufficient to equate supply and demand. For example, if there is excess demand for current goods, this excess demand may not be eliminated by increases in the current price level which are accompanied by equally large increases in the future price level. In a trading economy the effect of the price increases is to reduce the wealth of the traders toward the endowment point ($w(1)$, $w(2)$) in a two period model. Suppose there is only one good, which is perishable, and money is the only store of value. Then if the marginal utility of the current endowment exceeds the marginal utility of the second period endowment for all traders, the price of the good cannot rise high enough to reduce current demand to the current endowment. The same dilemma may arise when the Hicksian elasticity of expectations is equal to one, even though expected prices do not equal current prices.

Grandmont considers a model of this type where trading in futures contracts is excluded so that point expectations do not cause difficulties. It is a trading economy in which consumers receive an endowment of perishable goods in each period of their lives and an initial money stock in the first period. In the current period they maximize a utility function of consumption over the remaining periods of life (assuming the life span to be known) subject to budget constraints of the form $p_t x_t + m_t = p_t w_t + m_t - 1$,where future prices p_t are equal to functions ψ_t of present prices $p1$. He assumes

(1) The utility function $u_h\left(x_1, \ldots, x_{n(h)}\right)$ is continuous, increasing, and strictly quasi- concave for every h.
(2) The endowments w_t^h are positive for all h and t, $1 \leq t \leq n(h)$.
(3) Total money stock $M = \Sigma_h m_h$ is positive.

He then proves that the temporary monetary equilibrium exists, that is, money prices are well defined, if every agent's price expectations ψ_t^h are continuous and, for at least one agent, who will be living in the next period and who has a positive money stock, price expectations are bounded away from 0 and ∞. In Grandmont's opinion this result leaves the existence of temporary equilibrium 'somewhat problematic'.

However, it seems quite inappropriate to deal with a money which has no role to play in facilitating transactions. Grandmont and Younes (1972) have studied general equilibrium in a model similar to the model just described except that lifetimes are taken to be infinite and utility functions are separable by time period, that is $u^h(x_1, \ldots) = \sum_{t=1}^{\infty} \delta^t u^h(t)$ for $0 < \delta < 1$. Also money is now assigned a role in transactions, that is, only part of the proceeds of sales in the current period can be used to finance purchases in this period. Thus in each period there is both a budgetary constraint, as before and, in addition, a liquidity constraint, which may be written in simplest form as $p_t(x_t - w_t)^+ \leq m_t + kp_t(x_t - w_t)^-$, where for any vector z we write $z_i^+ = \max(z_i, 0)$ and $z_i^- = \max(z_i, 0)$, and $0 < k < 1$. Thus the fraction k of receipts from sales can be used to buy goods in the current period. This fraction could be allowed to vary by consumer and by good. The constraint on purchases is entirely in the spirit of Walras. It is an explicit modelling of a need for liquidity that he left implicit in his account.

In order to prove that a monetary equilibrium exists an assumption to bound expected prices is made which is very similar to the previous assumption for this purpose, and also very similar to the assumption made by Green to obtain existence of temporary equilibrium in a non-monetary economy with futures trading. The assumption is that the set of expected prices, over a finite planning horizon, that result from all possible choices of current prices, which are assumed positive, lie in a compact subset of the set of positive future prices. Then if all consumers have continuous

expectations which satisfy this assumption, and the assumptions of the previous model are also met, there will exist a temporary equilibrium in this case also. Indeed, the case $k = 1$, where the liquidity motive is lacking, can be allowed.

In the second model where money has a transactions role expectations are described as depending on past prices as well as current prices, which leads inelastic price expectations to be more plausible. It also gives plausibility to correct foresight in states of stationary equilibrium over sequences of periods. Grandmont and Younes (1973) prove that the stationary equilibria of the model are not Pareto optimal.

However, they can be made Pareto optimal by use of a lump sum tax to reduce the quantity of money by a factor equal to the discount factor for utility. It is then proved that a continuum of such equilibria exists to sustain any Pareto optimal allocation, since the price level falls by the same factor, and it is not worthwhile to reduce a money stock, even if it is in excess of transaction requirements. Moreover, if the tax rate is set slightly too high, the consumer will always wish to increase his real balances and no stationary equilibrium will exist. Grandmont and Younes are not able to prove that an exact stationary equilibrium exists for a fixed money stock, although a near equilibrium exists if the discount factor is near1.

In addition to proofs of existence and non-optimality for monetary equilibria, Grandmont and Younes show that the quantity theory holds between stationary equilibria, that is, if p and m_h, $h = 1,...,m$, provide a stationary equilibrium, then λ_p and λ_{mh} also provide one. This is the conclusion of Walras as well. On the other hand, the stationary equilibria of a monetary economy will differ from the stationary equilibria of a barter economy unless $\delta = 1$. This is apparent from the fact that the barter economy's equilibria are Pareto optimal and the monetary economy's equilibria are not, unless $\delta = 1$. Thus the simple 'classical dichotomy' does not hold.

Equilibrium Over Time

In addition to temporary equilibrium Hicks considered the possibility of equilibrium over time, in the sense that the expectations held by traders in one market about prices on future markets are realized when those future markets are held. However, when there is uncertainty it is not clear what is meant by the realization of expectations. If expectations take the form of a non-atomic probability measure over future prices, any vector of prices within the support of the measure is as likely as any other, that is, it has zero probability. Nor does the Hicksian trick of replacing the probability distribution by a representative price, depending on the trader, avoid the difficulty, since the representative price is not typically a statistic of the price distribution, such as the mean or the mode. Thus even if all traders held the same expectations in the sense of a probability distribution for prices, they would not have the same representative prices except by the chance that their circumstances and their risk preferences also coincide.

A way to resolve this dilemma was provided by Radner (1972). His solution is a type of perfect foresight. All traders hold the same point expectations for prices with certainty, contingent on the event in which the market is held. Only a finite number of dates are allowed and only a finite number of events may occur in each. From the viewpoint of a given market the relevant elementary events are the possible sequences of states of nature that may occur up to the horizon. For any such sequence the traders expect correctly a corresponding sequence of prices. This does not lead to a grand initial market in which all future exchanges are arranged because the set of forward commitments which are actually available in the market is a small subset of all those associated with future events. For example, it may be that most commodities are traded for sure delivery and only one commodity (money or the numeraire) is traded on a contingent basis (insurance). It should be noted that this construction does not depend on any agreement between traders on the probabilities of the alternative events. Thus the expectation functions which were introduced in the discussion of temporary equilibrium would not be likely to be the same for different traders.

In this setting the trader plans a sequence of consumptions contingent on the events in which

they occur and also a sequence of trades on the markets which are open. Spot markets are open for all commodities at all dates but only a small subset of the possible markets in forward contracts may be open at any particular date. In any case since the number of dates and states of nature and thus of elementary events is finite, only finitely many prices will arise.

Let X_h be the consumption set of the hth consumer. Let M be the set of elementary date-events pairs. A consumption-trade plan for the hth consumer is a pair (x^h, z^h) where x^h_m is the consumption planned for $m \in M$ and z^h_m is the trade planned for $m \in M$. Let $\Gamma^h(p)$ be the set of feasible plans for h, given prices p. In particular, (x^h,z^h) in $\Gamma_h(p)$ implies that consumption x^h_m plus net deliveries z^h_m due at m are not greater than resource endowments w^h_m for each m and the budget constraint $p_m z^h_m$ holds at each $m \in M$.

Let $\gamma^h(p)$ be the set of plans in $\Gamma_h(p)$ which are optimal for h. An equilibrium of plans and price expectations (including current prices which are known) is given by plans (x^h,z^h) and expected prices p such that (x^h,z^h) is in $\gamma^h(p)$ for each h, that is, the plans are preferred at the expected prices, and the sum $\sum_h z^h_m$ of commitments at each m is non-negative, and the value of commitments $p_m \sum_h z^h_m = 0$ at each m, that is, Walras's Law holds. In such a purely trading economy for perishable goods with a finite set of dates and events and under assumptions of the usual kind on preferences, and positive endowments which lie in the interior of consumption sets, Radner proves that an equilibrium exists.

It is not difficult to bring production into this setting if firms are introduced with fixed production plans and with shares which are traded on a stock exchange.

The ownership of a share of a firm can be equated to the ownership of a share of its output, including the end of the period capital stock. The output of a firm at any date would depend on the event, and the function relating this output to the events would be known by traders, just as future prices of goods are known, contingent on events. Now, in addition to goods prices, share prices are foreseen in each event at each date with certainty. As before the number of dates and events is finite.

A feature of this model not present in the trading model is that consumers do not own the resources of the firms as individual goods but as proportions of the batch of goods that firms hold. The consumer can buy and sell goods forward by means of long and short positions in the stock market but the trade he arranges by these means for one event at the next date determines his trade for all other events at that date. Thus spot markets still may offer useful alternatives, quite aside from the practical difficulties of physically dissolving the firm. Of course, given the presence of spot markets, dissolution of the firms is not needed if the value of the firm equals or exceeds the value of its resources.

If one tries to go further to specify how the production and trading plans are arrived at, a major problem arises of setting the objectives of the firm. Hicks solves this problem by assuming that the production plan chosen would have the maximum discounted value among those available. This value could be calculated since expectations were single-valued and interest rate, actual or expected, could be used in arriving at present values. Moreover, firms were treated like single proprietorships. In the modern literature firms have sometimes been assigned utility functions defined on the streams of profits. Another suggestion is to suppose that the firm adopts the plan that maximizes the value of its shares on the stock market. This would seem to be the approach most in accord with other parts of general equilibrium theory.

However, it encounters the difficulty that the judgement of the management and the judgement of the market on the probability of different events may not coincide. If this difference of judgement exists, the market solution would be for the firm to be purchased through a takeover by those who value its potential most highly and the management displaced. Markets which work in this way would correspond quite well to the original Walrasian model.

Various results on the existence of a general equilibrium have been reached with special models of production by firms. One theorem of Radner extends the existence of an equilibrium of plans and price expectations to this context. His assumptions are:

(1) Consumers satisfy the usual conditions on convexity, non-satiation, and positive endowments.
(2) Consumers own the shares of firms and each consumer owns shares in every firm.
(3) Producers have closed, convex production sets with free disposal. The total production set satisfies the condition that the negative of a producible vector of commodities is not producible.
(4) Each firm has a continuous, strictly concave utility function on profit streams.

With these assumptions he does not achieve a full existence theorem because the model is not well adapted to handle the entry and exit of firms. What may happen is that some firms show an excess supply of shares in some events and dates. Then since the firms are treated like partnerships with unlimited liability, negative share prices might be justified at this point. In any case the questions of entry and exit of firms is one that the Arrow–Debreu model also fails to deal with. The theorem proved by Radner only finds a 'pseudo-equilibrium' where the value of total excess supply (of shares) is minimized.

In the foregoing discussion it has been assumed that only a subset, possibly small, of the potential Arrow–Debreu markets is open. It is possible to justify the selection of markets which are open by postulating costs for carrying out transactions. If the markets which are open are given, the previous equilibria may be supported by assigning infinite transactions costs to the lost markets and zero costs to the open ones. Otherwise the open markets will be endogenous to the general equilibrium. In the analysis of markets with transaction activities which consume resources the same convexity or linearity assumptions have been used as for the production technology. Then it is not difficult to prove existence of equilibrium under assumptions of the usual sort.

Rational Expectations

It has been implicitly assumed in the preceding discussion of temporary equilibrium that the traders have the same information available. If this is not the case the complication arises that the equilibrium price may convey information. For example, in the market for umbrellas if some traders have the benefit of weather forecasts and some do not, a high price based on the demand of informed traders will signal to uninformed traders that rain is expected. Then all traders are informed and an equilibrium price must be consistent with fully informed demand.

A difficulty arises if it happens that the utilities of consumers depend on events in contrary ways, that is, uninformed consumers use umbrellas to ward off sun and informed traders to ward off rain. Then price will be higher if rainy weather is expected by informed traders but if uninformed traders perceive this and become informed, the high price may not appear and a fully informed market may not show a price difference depending on the weather forecast. But then no information is transmitted so the weather forecast cannot be read out of equilibrium prices. The conclusion is that no equilibrium is possible. However, the result requires an exact balance in the effects of rain and sun on the two sets of traders, so it is unlikely to hold. More robust examples of nonexistence were given by Green (1977) and Kreps (1977). The idea of the discontinuity was first proposed by Radner (1967).

A rational expectations equilibrium is said to exist if there is a function φ mapping states of the world into equilibrium prices which is invertible, that is, φ^{-1} exists, mapping prices, from a normalized set, into states of the world. It is clear that such a function will exist if the equilibrium price which appears when all traders are fully informed is uniquely determined by the elementary event, and the relation is one to one. It is also clear, given a finite set of elementary events, that the correspondence of prices to elementary events will be one to one in all but exceptional cases. Then the equilibrium is said to be revealing. But the price function of a revealing full information equilibrium is a price function that provides a rational expectations equilibrium. This observation is due to Grossman (1981).

The situation is more complicated when the possibility is recognized that spending resources

will allow more information to be gathered. The information that is disseminated free of charge by prices will discourage the use of resources to gather information and thus prevent the attainment of a Pareto optimum. In welfare terms a suboptimal amount of resources will be devoted to information activities.

An Infinite Horizon

In the Arrow–Debreu model of general equilibrium there are a finite number of periods, a finite number of locations, a finite number of events, and a finite number of commodity types, so the number of distinct goods when all these grounds for distinguishing goods have been recognized is still finite. The principal objection to the restriction to a finite number of goods is that it requires a finite horizon and there is no natural way to choose the final period. Moreover, since there will be terminal stocks in the final period there is no natural way to value them without contemplating future periods in which they will be used. The finiteness of the number of locations and commodity types is achieved by making a discrete approximation to a continuum, and perhaps the finiteness of the number of states of nature can also be viewed in this light. But in the case of time, a discrete approximation by periods still leaves a denumerable infinity of dates.

There are two principal models in which an infinite number of goods appear. In one model there is a finite number of infinitely lived consumers. Such a consumer may be considered to represent a series of descendants stretching into the indefinite future, so that consumers alive in the present period have an interest in the goods of all periods. The other model has an infinite number of consumers, but only a finite number of them are alive in any period. This model is called the overlapping generations model. It was first proposed and explicitly analysed by Samuelson (1958).

A model of general competitive equilibrium with a finite number of consumers and an infinite number of commodities was first presented in rigorous form by Peleg and Yaari (1970). They assumed the number of commodities to be denumerable.

This is a basic case since a noncompact but separable commodity space can be approximated arbitrarily closely with a denumerable set of commodities in the same sense that a compact commodity space can be approximated by a finite set of commodities. This assumes that a sensible neighbourhood system can be defined in the commodity space, as Debreu does for the dimensions of location and time with places and periods.

Peleg and Yaari present a trading model without production. The commodity space s is the space of all real sequences. In order to discuss continuity the space must be given a topology, in this case, the product topology. Thus a sequence of points converges if it converges in every coordinate, that is, $x^s \to x$, $s=1,2,...$, if $x^s(i) \to x(i)$ for $i = 0,1,...$. The space is presented as a sequence of real numbers but by grouping terms it may equally well represent a sequence of vectors, for example, commodity bundles occurring in successive time periods. The hth trader has an initial stock w_h, where $w_h \in s$ and a preference relation $\succsim_h$, which is reflexive, transitive, and complete on $s+$, the set of non-negative sequences. Strict preferences $\succ_h$ is defined by $x \succsim_h y$ if $x \succsim_h y$ and not $y \succsim_h x$.

Peleg and Yaari prove an existence theorem for this economy on the following assumptions.

(1) Desirability. If $x \geq y$, then $x \succsim_h y$.
(2) Strong convexity. If $x \neq y$ and $x \succsim_h y$, then $\alpha x + (1 - \alpha) y \succ_h y$ for $0 < \alpha < 1$
(3) Continuity. The two sets $\{y | y \succsim_h x\}$ and $\{y | y \precsim_h y\}$ are closed.
(4) Positivity of total supply. Let $w = \sum_{h=1}^{m} w_h$. Then $w > 0$.

A price system is a real sequence $\pi > 0$ which satisfies $\sum_{i=0}^{\infty} \pi(i) w_h(i) < \infty$, that is, the value of the initial bundles is finite. This implies that $\pi(i)$ $w_h(i)$ converges to 0 as $i \to \infty$. A competitive equilibrium is given by $(x_1,...,...x_m;...\pi)$ such that π is a price system, $\sum_{i=0}^{\infty} \pi(i) x_h(i) \leq \sum_{i=0}^{\infty} \pi(i)$ $w_h(i)$, for each h, and $\sum_{i=0}^{\infty} \pi(i) x(i) \leq \sum_{i=0}^{\infty} \pi(i)$

$w_h(i)$ implies $x_h \succsim_h x$. Peleg and Yaari prove that a competitive equilibrium exists.

It is clear from their discussions, and it has become even clearer in subsequent work, that the use of a topology such that, in the context of an infinite horizon interpretation of the model, impatience is implied by continuity of preferences is the crucial assumption for a proof of existence. That the product topology implies impatience may be seen in the following way. If $x \succsim hy$ then by continuity there is a neighbourhood U of x such that $z \in U$ implies $z \succ_h y$. However, a neighbourhood U is defined by $|z(i) - x(i)| \langle \in \rangle$ 0 for a finite number of coordinates where the remaining coordinates are free. Thus given $y \succsim_h x$, there must exist $N > 0$ such that $z(i) = x(i)$ for $i \leq N$ and $z(i)=0$ for $i > N$, and $z \succ_h y$. These conditions are met if the preference order is representable by a separable utility function which is the sum of period wise utilities discounted back to the present at a constant rate per period, and these utilities are continuous and uniformly bounded. Such a utility function is a common way of expressing impatience.

A model of general competitive equilibrium which allows for production where there is an infinite number of commodities was first presented in a rigorous form by Bewley (1972). A preference relation $\succsim h$ is assumed for each consumer as in Peleg and Yaari. We will describe Bewleys model for the case of a sequence of periods with an infinite horizon where $N1$ is the finite set of commodities available in the tth period. Then the set of all commodities is

$$M = \bigcup_{t=1}^{\infty} N_t.$$

It is assumed that $M = M_c \cup M_p$ where M_c and M_p are disjoint and M_c contains the consumption goods. Bewley confines attention to the commodity space l_∞ of bounded sequences of real numbers. Let $K_c = \{x \in l_\infty | \; x(i) = 0$ for not $i \in M_c$, $x(i) \geq 0$ for $I \in M_c\}$, and similarly for Kp. Let $\tilde{K}_c$ have the same definition as Kc except that $x(i) > \varepsilon > 0$ for all $i \in Mc$ for some given ε. Bewley's existence theorem holds for a weaker notion of continuity than that of component wise convergence, but we will stay with the definition used by Peleg and Yaari for the sake of simplicity.

Then the assumptions on the consumer sector are

(1) The consumption sets $X_h = K_c - w_h$ where w_h is the endowment of the hth consumer.
(2) The sets $\{y | y \succsim_h x\}$ and $\{y | x \succ = y\}$ are closed. Also $\{y | y \succsim_h x\}$ is convex.
(3) Mc is not empty and for each h, if $x \in Xh$ and $y \in \tilde{K}_c$, then $x + y \succsim_h x$.

The production sector is defined by means of production sets Yt which convert inputs belonging to Nt-1 into outputs belong to Nt. The $Y = \sum_{t=1}^{\infty} Y_t$.

The assumptions on the production sector are:

(4) Y is a convex closed cone with vertex at 0.
(5) If $w \in l_\infty$, then $Y + w \cap l_\infty$ is bounded.
(6) If $y \in Y$, then $y^n \in Y$ where $y_t^n = y_t$ for $t = 0,...,n$, and $y_t^n = 0$ for $t > n$.
(7) $-Kp \subset Y$

Assumption (4) means that each Yt is a linear activities model as Walras assumed. Assumption (5) excludes unbounded production from given inputs. Assumption (6) allows production to end at any time with free disposal of the final outputs. Assumption (7) allows free disposal of all goods other than consumption goods.

In addition there is one assumption which relates the consumption sector and the production sector.

(8) For each consumer h, there exists $\overline{x}_h \in X_h$ and $\overline{y}_h \in Y$ such that $\overline{y}_h(i) - \overline{x}_h(i) > \in > 0$ for all i and some $\in > 0$.

Assumption (8) protects consumer income in the sense that the consumer is not reduced to the subsistence level in equilibrium. That is to say, there are cheaper consumption bundles within this consumption set at equilibrium prices. An equilibrium is an allocation $(x_1...,...,.. .x_m,y)$ and a price sequence $\pi = (\pi(0), \pi(1),...)$ where $\pi(i)$ is non-negative for all i but different from zero for some i, which satisfy the conditions:

G

(I) $y \in Y$ and $\pi y = 0$, $\pi z \leq 0$ for all $z \in Y$. The profit condition.
(II) $x_h \in X_h$ and $\pi x_h = 0$, all h, and $z \succ_h x_h$ implies that $\pi z > \pi x_h$. The demand condition.
(III) $\sum_{h=1}^{m} x_h = y$. The balance condition.

On the basis of the assumptions Bewley is able to prove that an equilibrium exists where the price system $\pi \in l_1$, that is, $\sum_{i=0}^{\infty} \pi(i) < \infty$. This represents a generalization of the classical existence theorem in the form given by McKenzie to the case of denumerably many commodities, retaining the assumption of a finite number of consumers. The argument is stated in terms of an infinite horizon and a finite number of goods in each period, but the original theorem is more general and applies to the case of uncertainty with an infinite number of events as well as to models with a continuum of commodities. The continuum of commodities may arise from a variation in the physical properties of the goods and services.

Overlapping Generations

In the overlapping generations model of general equilibrium the number of consumers as well as the number of commodities is infinite. However, at any given time the number of both is finite. While the model with a finite number of infinitely lived consumers treats the consumers who are living as if their lives were extended into the indefinite future by the lives of their descendants, in the classical overlapping generations model bequests are neglected and each generation is assumed to be interested only in its own consumption.

The first rigorous analyses of an overlapping generations model in a general equilibrium setting were done by Balasko et al. (1980) and by Wilson (1981). They treat an exchange model in which all goods perish in each period, and each consumer receives an endowment in each period. They assume that each consumer lives for two periods. However, this assumption is not essential. What is essential is that lifetimes are finite in length and some of the people alive at any date have lifetimes which overlap the lifetimes of some people who are born later than they.

The formal model makes these assumptions.

(1) In each period t ($t = 1, 2,...$) there is an arbitrary, finite number of perishable commodities $n' \geq 1$.
(2) Each consumer $h = 1, 2,...$ lives for two periods. At the start of period t an arbitrary but finite number of consumers is born with indices $h \in G^t$.
(3) Consumption sets $X_h = R_+^{n(0)}$ for $h \in G0$ the consumers alive when the economy begins and $X_h = R_+^{n(t)} \times R_+^{n(t+1)}$ for $h \in G^t$, t ≥ 1. Write $x_h = x'_h$ for $h \in G^0$ and $x_h = (x_h(t), x_h(t+1))$ for $h \in G^t$.
(4) Each consumer has a utility function, $u_h(x(1))$ for $h \in G^0$ and $u_h(x(t), x(t+1))$ for $h \in G^t$. Utility functions u_h are continuous, quasi-concave, and without local maxima.
(5) Each consumer receives an endowment, $w_h = w_h(1)$ for $h \in G^0$ and $w_h = (w_h(t), w_h(t+1))$, for $h \in G'$. For each h, $w_h \geq 0$ and $w_h \neq 0$.
(6) The economy is inter-temporally irreducible. Let $I(t) = \{h |\ h \in G^s \text{ for } 0 \leq s \leq t\}$. Then there exists a sequence $t_\mu \to \infty$ with the following property. Given any allocation $x = (x_1, x_2,...)$ and $I_1(t\mu)$and $I_2(t\mu) \neq \varphi$, with $I_1(t_\mu) \cap I_2(t_\mu) = \varphi$, and $I_1(t_\mu) \cup I_2(t_\mu) = I_1(t_\mu)$, there exist $y_h \geq 0$ for $h \in I1(t_\mu)$ and $x'_h \geq 0$ for $h \in I_2(t\mu)$, such that $\sum_{h \in I_1(t_\mu)} y_i(t) = 0$ when $\sum_{h \in I_1(t_\mu)} w_{hi}(t) = 0$ for $1 \leq i \leq n^t$, $1 \leq t \leq t_\mu + 1$, and

$$\sum_{h \in I_2(t\mu)} x'_h \leq \sum_{h \in I_1(t\mu)} (w_h + y_h) + \sum_{h \in I_2(t\mu)} w_h.$$

Moreover, $u_h(x'_h) \geq u_h(x_h)$ for all $h \in I_2(t_\mu)$ with the strict inequality for some h.

Assumption (6) is the irreducibility assumption of McKenzie adapted to economies made up of the consumers born by the period $t\mu$. It says that it is always possible to increase the welfare of the second subgroup if the scale of the endowment of the first subgroup is increased.

Let $p = (p(1), p(2), \ldots)$ where $p(t) \in R^{n(t)}$. Then the pair (x, p) is a competitive equilibrium if

(I) For all h, $u_h(x_h)$ is maximal over all z_h such that

$$p(t)z_h(t) + p(t+1)z_h(t+1) \leq p(t)w_h(t) + p(t+1)w_h(t+1) \quad \text{if} \quad h \in G^t,\ t \geq 1, \text{and} \quad p(1)z_h(1) \leq p(1)w_h(1) \text{ if } h \in G^0, \text{where } z_h \geq 0.$$

(II) $\sum_h x_{hi}(t) \leq \sum_h w_{hi}(t)$ with equality if $p_i(t)>0$, where the summation is over $h \in G^{t-1} \cup G_t$, $1 \leq i \leq M(t)$ and $t \geq 1$.

Condition (I) is the usual demand condition and condition (II) is the balance condition. Balasko et al. (1980) prove that the six assumptions listed imply the existence of a competitive equilibrium. They show that the artificial assumptions on birthdates and lifetimes are irrelevant by a redefinition of the period. They also conjecture that the introduction of production and consumption sets of the usual classical type, which are closed, convex, and bounded below, would cause no major difficulties.

Wilson (1981) treats an economy which may contain both finite lived and infinite lived consumers and which may be specialized to either. He also allows intransitive preferences. He uses a somewhat simpler version of irreducibility and proves existence in an exchange economy where the number of goods in each period is finite in two circumstances (1) when the consumers are all finite lived and (2) when a finite subset of infinite lived consumers own a positive fraction of the endowment in all but a finite number of periods. If preferences are transitive and strictly convex, the competitive equilibrium is also Pareto optimal. Thus Wilson's results contain the theorems on existence of Bewley and Balasko, Shell, and Cass as special cases while also providing conditions in the model sufficient for Pareto optimality.

A striking difference between the competitive equilibria of economies where the number of consumers is finite, and the competitive equilibria of economies with overlapping generations and an infinite horizon, where the number of consumers is infinite, is that with perfect foresight the former equilibria are also Pareto optima while the latter need not be. This is the major point emphasized by Samuelson in his initial paper. The most general theorem proving that competitive equilibria are Pareto optima even when the number of commodities is infinite provided that the number of consumers is finite is due to Debreu (1954). Under some additional smoothness conditions on utility and boundedness conditions on prices and allocations Balasko and Shell (1980) prove that the allocation x of a competitive equilibrium is Pareto optimal if and only if $\sum_t (1/\| pt\|) = \infty$. This is a condition which had already been shown to characterize efficiency in neoclassical production economies by Cass. It is clear that $\lim inf (\| p_{t+1}\| / \| p_t\|) = r \leq 1$ implies that the condition for Pareto optimality is satisfied since the sums dominate $\Sigma t\ (1/r^t)$ which diverges. Intuitively, for a stationary economy if the interest rates are asymptotically non-negative, the competitive equilibria will be Pareto optimal, or if the economy is growing, if the interest rates exceed the growth rate, Pareto optimality follows.

Limitations of the Analysis

As mentioned in the beginning the claim of the theories described as general equilibrium theories to be 'general' is qualified by the set of conditions considered to be constant. Walras as well as most subsequent theorists classified the constant factors as tastes, technology, and resources, including population. However, all three of these categories have been treated by some economists as responding, in ways amenable to analysis, to market variables. These studies have usually been confined to a few variables and have usually been partial equilibrium in character, although the classical school of economists included population as a major variable in models of economic development. Their models are comprehensive but lack the market equilibrium analysis of the general equilibrium theories, whose inspiration appears to have been found in the marginal utility theory of consumer demand. Similarly, tastes

have sometimes been modelled to depend on past consumption or advertising, and technology has been modelled to depend on research and development spending and on the rewards to innovation. Also natural resources, in terms of resources known to exist, are often treated as responding to prices.

From this perspective general equilibrium theory is a partial theory of economic affairs with a special set of *ceteris paribus* assumptions. The variables which are left free are chosen because they lend themselves to a particularly elegant theory in terms of consumer demand under budget constraints and producer supplies with profit conditions where these constraints and conditions are established by prices equating demand and supply. This was the vision of Walras, perhaps guided by the theory of static equilibrium of mechanical forces which he found in Poinsot.

Another direction of abstraction in general equilibrium theory in its classic expressions has been to ignore the effects of processes which do not pass through the market. In particular each consuming unit is described as interested only in its own consumption in the theory of Pareto optimality and as uninfluenced in its choices by the choices made by other households. Similarly, the production possibilities of one firm or process are treated as independent of the productive activities of other firms.

Some attempts have been made to incorporate these effects in the general equilibrium models but not with complete success. In particular there is not a good theory of existence when consumer possibility sets or production sets are affected by levels of consumption and production.

The convexity assumptions which have appeared in general equilibrium models from the time of Walras are often not good approximations of reality though they are depended on for many of theorems of the subject, such as the theorems on existence and Pareto optimality. However, there is a theory of approximate equilibria and of limiting results as the size of the market increases relative to the participants which does something to bridge the gap between theory and fact.

Finally, the assumption that the market participants take prices as independent of their actions fails to describe many markets, and describes very few exactly.

Nonetheless, this assumption may be useful for a theory that embraces all markets, whose special features cannot be described in detail. It may, that is, give a good approximation to the working of the economy as a whole. Also it is useful for its implications for optimality, a point which was perceived, albeit through a glass darkly, by Walras. The proper notion was later found by Pareto.

Just as the model does not accommodate monopoly easily, government does not fit in well. A chief difficulty arises from its compulsory features which allow it to extract resources by force rather than by voluntary agreement. Government is not easily described either as a producer selling services, or as a voluntary organization performing acts of collective consumption, though in ways it resembles both.

Voluntary societies also do not fit perfectly in the scheme of producers and households though the disparity is less, since they must meet their expenses from contributions by the membership who will not contribute unless the services of the society to them are worth the dues they pay.

Properties of General Equilibrium

Walras set the major objectives of general equilibrium theory as they have remained ever since. First, it was necessary to prove in any model of general equilibrium that the equilibrium exists. Then its optimality properties should be demonstrated. Next it should be shown how the equilibrium would be attained, that is, the stability of the equilibrium and its uniqueness should be studied. Finally, it should be shown how the equilibrium will change when conditions of demand, technology, or resources are varied, the subject now called comparative statics. He contributed to all these lines of research.

Walras's arguments for existence are not conclusive but he did contribute a basic principle, that the model should be neither underdetermined nor over determined. That is, the number of independent equations to be satisfied and the number of

variables to be determined should be equal. Some critics saw right away that this equality did not ensure a meaningful solution to the equation system, for example, that the solution to such an equation system is not guaranteed to be real.

The question was not taken up seriously until the 1930s and the first rigorous treatment was given by Wald (1935, 1936). Then in the 1950s more complete solutions on neo-classical assumptions were found by Arrow and Debreu (1954), McKenzie (1954), and Nikaido (1956).

In the discussion of models of general equilibrium that have been given above, the first requirement has been a set of assumptions from which existence could be inferred. This approach to the subject was begun in the papers of Wald and von Neumann, presented to the colloquium of Karl Menger (mathematician and son of Carl Menger, the neoclassical economist) in Vienna in the 1930s.

The optimality that Walras claimed for competitive equilibrium, under conditions of certainty, except for insurable risk, did not seem to go beyond individual maximization of utility in face of an equilibrium price system. However, Pareto gave a genuinely social definition that the allocation of goods and services in a competitive equilibrium is such that no reallocation is possible with some consumer better off unless some consumer is made worse off. In fact, Walras seemed to be groping for the same definition and his arguments may be slightly extended to establish Pareto's proposition.

As noticed in the earlier discussion of markets with certainty, Pareto optimality is implied by maximization of preference under budget constraints and von Neumann's law, or maximization of profit given the technology. The former implies that an allocation which improves one consumer's position and harms none must, given local non-satiation for all consumers, be more valuable at equilibrium prices while the latter implies that no more valuable allocation is achievable. This argument depends on the finiteness of the value of the goods in the economy. Otherwise the impossibility of a more valuable allocation is not meaningful. Thus when the horizon is infinite and the discount factor is too large, for example, equal to 1 if the economy is stationary, or in general greater than or equal to the reciprocal of the growth rate, Pareto optimality may fail in competitive equilibrium, as Samuelson showed. Also there is no reason to expect Pareto optimality, in an exact sense, when some markets are missing, a very likely eventuality when there is uncertainty and goods must be traded on every possible contingency to provide complete markets.

A second theorem on Pareto optimality asserts that any Pareto optimum can be realized as a competitive equilibrium. This theorem requires assumptions which are similar to those leading to existence, in particular, assumptions providing local non- satiation for some consumers and convexity of the preferred sets and the feasible set. Moreover, when the number of goods is infinite as in the case of an infinite horizon an additional condition is needed to give the existence of the prices. This condition may be that the sum of consumers' preferred sets has an interior or that the production set has an interior. In the case of the product topology and free disposal by consumers the preferred sets will have interiors if the period wise utility functions are continuous and bounded (see Debreu 1954). Finally it was shown by Arrow (1953) that in order for the Pareto optimal allocation to maximize preference over the budget set rather than only to minimize the cost of achieving a given preference level, it is useful to assume that xi, the consumption set of the ith consumer, contains a point which is cheaper than the allocation he receives, for $i = 1,...,m$.

The stability theory for general equilibrium has been largely devoted to the stability of the Walrasian tâtonnement, or process of groping for equilibrium prices through a process of price revision according to excess demand. That is, prices rise or fall depending on whether excess demand is positive or negative. In the tâtonnement there is no trading until equilibrium prices have been reached. The most convincing theorems concern local stability and the dominant assumption leading to local stability is that the market excess demand function satisfies the weak axiom of revealed preference between the equilibrium price and any other price in a sufficiently small neighbourhood of the equilibrium price. That is, if

$\overline{p}$ is an equilibrium price and e is the excess demand function, $p \cdot e(\overline{p}) - p \cdot e(p) \leq 0$ implies $\overline{p} \cdot e(\overline{p}) - \overline{p} \cdot e(p) < 0$. Since $\overline{p}$ is an equilibrium price, $e(\overline{p}) = 0$, and $p \cdot e(p) = 0$ by Walras's Law. Therefore, the condition holds and we may conclude that $\overline{p} \cdot e(p) > 0$.

The weak axiom for the market may be expected to hold if the net income effect of price changes is small.

Consider the price revision process given by $\mathrm{d}p_i/\mathrm{d}t = \dot{p}_i = e_i(p), \ i = 1, \ldots, n-1$, where the nth good is numeraire so $\dot{p}_n \equiv 0$. Then consider the function $|p(t) - \overline{p}|^2$, the square of the distance from the equilibrium price vector to the price vector at time t. We derive

$$\mathrm{d}/\mathrm{d}t(|p(t) - \overline{p}|)^2 = 2\sum_1^n (p_i - \overline{p}_i)\dot{p}_i$$
$$= 2\sum_1^n (p_i - \overline{p}_i)e_i < 0,$$

using the weak axiom of revealed preference and Walras's Law. Thus the distance of $p(t)$ from $\overline{p}$ constantly falls, or $p(t) \to \overline{p}$ as $t \to \infty$. Since locally the rate of price change can be equated to excess demand for any continuous tâtonnement by choice of units, this is a general argument. Since the assumption of gross substitutes ($e_{ij} < 0$ for $i \neq j$ and $e_{ij} = \partial e_i(p)/\partial p_j$) implies the weak axiom, and the assumption of a negative definite Jacobian $[e_{ij}]$, $i, j = 1, \ldots, n-1$, at equilibrium is equivalent to the weak axiom locally, the weak axiom is a dominant condition for local stability. All global stability results are very special and relatively unconvincing.

A rigorous treatment of the stability problem for the tâtonnement was given by Arrow and Hurwicz (1958) and Arrow et al. (1959). A stability theory which allows for trading was given by Hahn and Negishi (1962). These theories do not allow for speculative trading although profitable arbitrage opportunities would be likely to exist for any speculator who correctly inferred what the price revision process was. The stability of the tâtonnement was conjectured by Walras, to be the normal case for economies with many goods and essentially correct arguments were given by Walras for the case of exchange economies with two goods. He recognized and illustrated the case of locally unstable equilibria in the two goods case.

Finally, as Walras saw, it may be possible through a general equilibrium analysis to determine the effect of changes in the exogenous factors, resources, technology, or tastes, on the economic variables in equilibrium. This is analogous to the effect of a change in the constraints on the equilibrium of mechanical forces, an analogy with which Walras would have been familiar from the book of Poinsot. In the case of the exchange of two commodities Walras derives some simple and correct results for comparative statics just as he does for stability. He observes that an increase in the marginal utility of a good or a reduction in its supply will raise its price. In drawing this inference from his demand and offer curves he confines himself to stable equilibria as the only equilibria of interest.

Hicks used the comparative static result of Walras in a market with many goods to define stability of equilibrium. Samuelson (1947) pointed out that stability of equilibrium, where stability is given a dynamic interpretation as in a continuous tâtonnement, may imply comparative static results as a general principle. However, the straightforward generalization of Walras is the use of conditions which are sufficient to imply stability as a basis for deriving theorems on comparative statics. The most interesting theorem may be that derived from the revealed preference assumption at equilibrium.

Suppose that $e(\overline{p}) = 0$ but excess demand changes so that the new excess demand function $e'_n(\overline{p}) = e_i(\overline{p})$ for $i \neq 1$ or n and $e'_1(\overline{p}) = \delta_1 > 0$. while $e'_n(\overline{p}) = \delta_n > 0$. Let n be numeraire. This change can be arranged by taking δ_n of the nth good from some holder and compensating him with $\delta_1 = \delta_n/\overline{p}_1$ of the first good.

Suppose that the new equilibrium price is p, or $e'(p) = 0$. By Walras's Law $\overline{p} \cdot e'(\overline{p}) = 0$ and by the assumption of revealed preference $p.\ e'(\overline{p}) > 0$. Thus $(\overline{p} - p) \cdot e'(\overline{p}) < 0$, or $(\overline{p}_1 - \overline{p}_1)\delta_1 < 0$, or $\overline{p}_1 - p_1 > 0$. Any good falls in price when the excess demand for the numeraire rises at the expense of that good (see Allingham 1975).

A type of stability has been proved for competitive equilibrium over time which concerns the path of equilibrium prices over real time rather than the path of disequilibrium prices over virtual time, that is, the time of the tâtonnement. It was shown by Negishi (1960) that there is a social welfare function associated with a competitive equilibrium which is maximized in the equilibrium over feasible allocations. Suppose each consumer has a concave utility function which is given by a discounted sum of periodwise utilities. Then the social welfare function which is maximized is also a discounted sum of periodwise utilities equal to a weighted sum of the individual utilities. Then using results from turnpike theory for optimal capital accumulation it has been shown by Bewley (1982) that the competitive equilibrium allocations converge over time to the allocations of a stationary competitive equilibrium whose capital stocks and allocations are the same as those of the unique optimal stationary path of capital accumulation given the social welfare function. The utility functions and the production functions are assumed to be strictly concave and the discount factors are the same for all consumers and sufficiently near 1. However, these conditions may be relaxed.

Comparative static and comparative dynamic results have been derived from stability conditions in the context of optimal capital accumulation, which is equivalent to competitive equilibrium over time with a representative consumer. We may say that an optimal stationary path of capital is regular if an increase in the discount factor implies an increase in the value of capital stocks at initial prices.

Then there are sufficient conditions for local stability of the optimal stationary path which imply that the path is regular. Similar dynamic results may be achieved for non-stationary paths as well (see Araujo and Scheinkman 1979). It may be possible to extend these results to Bewley-type economies.

See Also

- ▶ Arrow–Debreu Model of General Equilibrium
- ▶ Existence of General Equilibrium
- ▶ Mathematics and Economics
- ▶ Overlapping Generations Model of General Equilibrium
- ▶ Uncertainty and General Equilibrium

Bibliography

Allingham, M. 1975. *General equilibrium*. New York: Wiley.

de Araujo, A.P., and J.A. Scheinkman. 1979. Notes on comparative dynamics. In *General equilibrium, growth, and trade*, ed. J.R. Green and J.A. Scheinkman. New York: Academic.

Arrow, K.J. 1951. An extension of the basic theorems of classical welfare economics. In *Proceedings of the second Berkeley symposium*, ed. J. Neyman. Berkeley: University of California Press.

Arrow, K.J. 1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. *Econométrie*, Paris: Centre National de la Recherche Scientifique. Trans. as 'The role of securities in the optimal allocation of risk-bearing', *Review of Economic Studies* 31 (1964): 91–96.

Arrow, K.J., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.

Arrow, K.J., and L. Hurwicz. 1958. On the stability of the competitive equilibrium I. *Econometrica* 26: 522–552.

Arrow, K.J., H.D. Block, and L. Hurwicz. 1959. On the stability of the competitive equilibrium II. *Econometrica* 27: 82–109.

Balasko, Y., and K. Shell. 1980. The overlapping generations model, I: The case of pure exchange without money. *Journal of Economic Theory* 23: 281–306.

Balasko, Y., D. Cass, and K. Shell. 1980. Existence of competitive equilibrium in a general overlapping generations model. *Journal of Economic Theory* 23: 307–322.

Bewley, T.F. 1972. Existence of equilibria in economies with infinitely many commodities. *Journal of Economic Theory* 4: 514–540.

Bewley, T.F. 1982. An integration of equilibrium theory and turnpike theory. *Journal of Mathematical Economics* 10: 233–268.

Cassel, G. 1918. *Theoretische Sozialökonomie*. 5th German edn, trans. as *The theory of social economy*. New York: Harcourt Brace, 1932.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie des richesses*. Paris: Hachette. Trans. as *Researches into the mathematical principles of the theory of wealth*. New York: Kelley, 1960.

Debreu, G. 1954. Valuation equilibrium and Pareto optimum. *Proceedings of the National Academy of Sciences* 40: 588–592.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Debreu, G. 1962. New concepts and techniques for equilibrium analysis. *International Economic Review* 3: 257–273.

G

Gossen, H. 1854. *Entwicklung der Gesetze des menschlichen Verkehrs*. 3rd ed, 1927. Berlin: Prager.
Grandmont, J.M. 1973. On the efficiency of a monetary equilibrium. *Review of Economic Studies* 40: 149–165.
Grandmont, J.M. 1977. Temporary general equilibrium theory. *Econometrica* 45: 535–572.
Grandmont, J.M. 1983. *Money and value*. New York: Cambridge University Press.
Grandmont, J.M., and Y. Younes. 1972. On the role of money and the existence of a monetary equilibrium. *Review of Economic Studies* 39: 355–372.
Green, J.R. 1973. Temporary general equilibrium in a sequential trading model with spot and future transactions. *Econometrica* 41: 1103–1123.
Green, J.R. 1977. The nonexistence of informational equilibria. *Review of Economic Studies* 44: 451–463.
Grossman, S.J. 1981. An introduction to the theory of rational expectations under asymmetric information. *Review of Economic Studies* 48: 541–560.
Hahn, F.H., and T. Negishi. 1962. A theorem on non-tâtonnement stability. *Econometrica* 30: 463–469.
Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.
Jevons, W.S. 1871. *The theory of political economy*. 5th ed. London: Macmillan. New York: Kelley and Millman, 1957.
Koopmans, T.C. 1951. Analysis of production as an efficient combination of activities. In *Activity analysis of production and allocation*, ed. T.C. Koopmans. New York: Wiley.
Kreps, D.M. 1977. A note on fulfilled expectations equilibria. *Journal of Economic Theory* 14: 32–43.
Marshall, A. 1890. *Principles of economics*. 8th ed, 1920. London: Macmillan.
McKenzie, L.W. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22: 147–161.
McKenzie, L.W. 1959. On the existence of general equilibrium for a competitive market. *Econometrica* 27: 54–71.
McKenzie, L.W. 1981. The classical theorem on existence of competitive equilibrium. *Econometrica* 49: 819–841.
Menger, C. 1871. *Grundsätze der Volkswirtschaflehre*. Vienna. Trans. *as Principles of economics*. Glencoe: Free Press, 1950.
Mill, J.S. 1848. *Principles of political economy*. London: Parker. New ed. London: Longmans, 1909.
Negishi, T. 1960. Welfare economics and existence of an equilibrium for a competitive economy. *Metroeconomica* 12: 92–97.
von Neumann, J. 1937. Über ein Ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes. *Ergebnisse eines Mathematischen Kolloquiums* 8: 73–83. *Trans. in Review of Economic Studies* 13, (1945): 1–9.
Nikaido, H. 1956. On the classical multilateral exchange problem. *Metroeconomica* 8: 135–145.
Novshek, W., and H. Sonnenschein. 1980. Small efficient scale as a foundation for Walrasian equilibrium. *Journal of Economic Theory* 22: 243–255.
Pareto, V. 1909. *Manuel d'économie politique*. Paris. Trans. from 1927 edn as *Manual of political economy*. New York: Kelley, 1971.
Peleg, B., and M.E. Yaari. 1970. Markets with countably many commodities. *International Economic Review* 11: 369–377.
Poinsot, L. 1803. *Eléments de statique*. 8th ed. Paris, 1842.
Radner, R. 1967. Equilibre des marchés à terme et au comptant en cas d'incertitude. *Cahiers d'Econométrie* 4: 35–52. Paris: CNRS.
Radner, R. 1972. Existence of equilibrium of plans, prices and price expectations in a sequence of markets. *Econometrica* 40: 289–303.
Samuelson, P.A. 1947. *Foundations of economic analysis*. Cambridge, MA: Harvard University Press.
Samuelson, P.A. 1958. An exact consumption – loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66: 467–482.
Scarf, H.F. (With T. Hansen) 1973. *The computation of economic equilibria*. New Haven: Yale University Press.
Smith, A. 1776. In *An inquiry into the nature and causes of the wealth of nations*, ed. E. Cannan, 5th ed. London: Methuen. 1906.
Wald, A. 1935. Über die eindeutige positive Losbarkeit der neuen Produktionsgleichungen. *Ergebnisse eines mathematischen Kolloquiums* 6: 12–20.
Wald, A. 1936. Über einige Gleichungssysteme der mathematischen Ökonomie. *Zeitschrift für Nationalökonomie* 7: 637–670. Trans. as 'On some systems of equations of mathematical economics', *Econometrica* 19, (1951): 368–403.
Walras, L. 1874–7. *Elements d'économie politique pure*. Lausanne: Corbaz. Trans. by W. Jaffé as *Elements of pure economics*. London: George Allen & Unwin, from the 1926 definitive edition, 1954.
Weintraub, E.R. 1983. On the existence of a competitive equilibrium: 1930–1954. *Journal of Economic Literature* 21: 1–39.
Wilson, C.A. 1981. Equilibrium in dynamic models with an infinity of agents. *Journal of Economic Theory* 24: 95–111.

General Equilibrium (New Developments)

William Zame

Abstract

General equilibrium theory is the theory of mass markets. The foundations of general equilibrium theory were laid in the late 19th and early 20th centuries by Walras and

Edgeworth. The modern formulation was conceived in the 1950s by Arrow, Debreu and McKenzie, who also established the fundamental results: existence of competitive equilibrium, Pareto optimality of equilibrium allocations (the First Welfare Theorem), and supportability of Pareto optimal allocations as equilibria with transfers (the Second Welfare Theorem). The ideas of general equilibrium theory are widely used in models of markets of all kinds, including in finance, international trade and macroeconomics.

Keywords

Adverse selection; Aggregate excess demand function; Arrow–Debreu model of general equilibrium; Asymmetric information; Bargaining; Commodity space; Competitive equilibrium; Convexity; Core convergence; Core equivalence; Default; Degree theory; Edgeworth, F. Y.; Efficient markets hypothesis; Equity premium; Existence of equilibrium; General equilibrium; Implicit function theorem; Incentive- compatible core; Incomplete markets; Kakutani fixed point theorem; Law of demand; Lipschitz functions; Market power; Moral hazard; Multiple equilibria; Perfect competition; Pooling; Private core; Private information; Rational expectations equilibrium; Revealed preference; Sard's theorem; Separation theorem; Tâtonnement; Transferable utility; Transversality conditions; Uniqueness of equilibrium; Walras's Law; Walrasian expectations equilibrium

JEL Classifications

D5

The fundamental ideas, results and applications of general equilibrium theory are well described in general equilibrium, which was originally written for the first (1987) edition of *The New Palgrave.* However, there has been a great deal of notable work since – too much to adequately survey in the limited space available here. The discussion addresses only a few topics on which research has been especially active:

- determinacy of equilibrium
- perfect competition and justification of the assumption of price-taking
- equilibration
- infinitely many commodities
- incomplete markets
- hidden information and hidden actions.

Determinacy

For many purposes, it is not enough to know simply that competitive equilibrium exists; we would like to know how equilibrium varies when the underlying parameters of the model vary. Such comparative statics analysis is simplest and most convincing when equilibrium is unique and depends nicely on the underlying parameters. However, it has been known for a long time that even some very simple economies admit multiple equilibria and that conditions on the primitives of an economy that guarantee uniqueness of equilibrium must necessarily be unpleasantly strong. For many purposes, however, it is enough to know that equilibria are locally unique, and locally depend nicely on underlying parameters.

Competitive equilibrium prices are the solutions of the system of equations asserting, for each good, that demand equals supply (equivalently, that aggregate excess demand is zero). In an economy with L consumption goods, this is a system of L equations with L unknowns; taking account of the price normalization and Walras's Law (aggregate expenditure equals aggregate income), this reduces to a system of $L - 1$ equations in $L - 1$ unknowns. Because the number of equations equals the number of unknowns, heuristic considerations (linear approximations, for example) suggest that local uniqueness might not be too much to hope for. However, local uniqueness does not *always* obtain: it is easy to exhibit simple Edgeworth box (two persons, two goods) exchange economies for which the set of equilibrium prices is a continuum.

Debreu (1970, 1972) showed that, if preferences are sufficiently smooth and indifference surfaces are not flat and do not intersect the boundary (for example, if preferences arise from utility functions that are twice continuously differentiable and differentiably strictly concave, and exhibit infinite marginal utility for consumption at zero levels of consumption), then almost all specifications of initial endowments lead to a finite number of equilibria, and those equilibria depend locally smoothly on endowments. Debreu's method was to first use the assumptions on preferences to show that the aggregate excess demand mapping is continuously differentiable, then to rely on Sard's theorem (which guarantees that for almost every $y \in Y$ every point in the inverse image $f^{-1}(y)$ is regular), and finally to appeal to the implicit function theorem.

Two limitations of Debreu's analysis are of particular note. The first is the assumption that indifference surfaces do not intersect the boundary. An implication of this assumption is that, at equilibrium, every agent consumes every commodity. This certainly seems false-to-fact, and this assumption would be objectionable in many applied models. Boundary consumptions create problems because they lead (almost necessarily) to an aggregate excess demand function that is not differentiable. Shannon (1994) extends Debreu's results, obtaining generic determinacy while accommodating boundary consumptions, by showing that the aggregate excess demand function, although not differentiable, is Lipschitz (that is, $|f(x) - f(y)| \leq C |x - y|$ for some constant C) and that the required implications of smooth analysis remain valid for Lipschitz functions. Blume and Zame (1993) follow a different approach, based on real algebraic geometry and particularly useful for applied models, treating only utility functions that are sufficiently smooth (roughly, piecewise real-analytic), but accommodating both boundary consumptions and utility functions that are not strictly concave.

A second, more subtle, limitation of Debreu's analysis is that the set of boundary endowments is of measure zero. Hence, saying that 'almost all specifications of initial endowments lead to a finite number of equilibria' says nothing at all about an environment in which some agents are not endowed with strictly positive amounts of all commodities. If it is true that most economic agents do not consume all goods, it is even more true that most economic agents are endowed with only a few goods – perhaps even with their own labour and nothing else. A more satisfactory specification would allow for the possibility that some agents' endowments of some goods are constrained to be zero, and to ask for determinacy for generic specifications of *other* goods. Surprisingly, Minehart (1997) finds that such specifications are compatible with robust indeterminacy. Mas-Colell (1985) and Anderson and Zame (2001) show that generic determinacy is restored if we interpret genericity in the sense of preferences (or utility functions) as well as endowments.

For economies with an infinite dimensional space of commodities, Debreu's arguments are typically inapplicable because the commodity space and the price space are different and demand functions are almost never continuously differentiable (Araujo 1987). However, the same outline can be applied, not to the aggregate excess demand mapping, but to the aggregate excess spending mapping. (Given a vector $\lambda = (\lambda_i)$ of utility weights for the agents, find the (unique) allocation (x_i) that maximizes the weighted sum $\Sigma\lambda_i\, u_i\,(x_i)$ of agent utilities and the price p that supports this allocation. The value of the excess spending map S at weights (λ_i) is the vector $S(\lambda) = (p \cdot x_i - p \cdot e_i)$ The price p and allocation (x_i) constitute a competitive equilibrium if $S(\lambda) = 0$.) Under the assumptions that utility is separable (across time or across states of the world) and that the underlying felicity functions satisfy Debreu's assumptions (at each date or in each state of the world), Kehoe and Levine (1985) show that the excess spending map is smooth and hence that equilibria are generically finite in number and depend nicely on parameters. In the general case, Shannon (1999) and Shannon and Zame (2002) identify conditions on utility functions implying that the excess spending map is Lipschitz; an application of Lipschitz analysis again yields generic determinacy.

Perfect Competition and Price-Taking

The definition of competitive equilibrium rests on the assumption of perfect competition; that is, that agents are price-takers. This assumption is clearly untenable if some agents are large, in the sense of controlling resources that are a significant fraction of the social total (although competitive equilibrium may serve as a useful benchmark even in such environments). A large and important literature attempts to understand when the assumption of price-taking behaviour is sensible.

One of the central themes in this literature seeks to justify price-taking behaviour by showing that cooperative outcomes are competitive – or almost competitive – when the population is large and agents are small. The largest portion of this literature is inspired by Edgeworth, who gave an informal argument that, for economies with two commodities and two consumers, replication shrinks the core of the economy (the set of feasible allocations on which no coalition can improve using only its own resources) to the set of competitive allocations. Debreu and Scarf (1963) give a formal statement of Edgeworth's assertion and show that it holds for economies with any (finite) number of commodities and consumers (assuming that consumers have strictly convex preferences). Formally: as the number of consumers grows the core coincides to the set of competitive allocations. Aumann (1964, 1966) constructs a formal limit model (with a continuum of consumers) and establishes (under quite general assumptions on preferences) that the core coincides with the set of competitive allocations.

Although these results are exceedingly elegant, the 'real' economy is finite and that replication and strict convexity of preferences are strong assumptions; hence neither Aumann's core equivalence theorem nor Debreu and Scarf's core convergence theorem applies directly to the 'real' economy. The results of Keiding (1974), Dierker (1975) and Anderson (1978) go a long way to removing this objection, showing that for *every* finite economy every core allocation can be approximately decentralized by some price, and that the deviation from exact decentralization (by some measures) is small provided the number of consumers is large and that no consumer is endowed with more than a small fraction of the social total of any good. For an excellent survey of the state of the art in the late 1980s, see Anderson (1993).

Although these (and related) results are usually accepted as cooperative justification for price-taking, more recent work reveals surprising subtleties.

1. Aumann's core equivalence theorem for continuum economies assumes only that preferences are locally non-satiated (not necessarily monotone). It had been widely assumed that convergence theorems for finite economies should obtain under the same assumption. (The results of Debreu and Scarf, Keiding, Dierker, and Anderson assume that preferences are strictly monotone.) However, Manelli (1991a, b) shows that core convergence may fail if preferences are not monotone, and Hara (2005) shows that further problems may arise if some commodities are bads.
2. Most of the work on core convergence treats only exchange economies. Xiong and Zheng (2005) show that the validity of core convergence for production economies depends in a subtle way on smoothness of preferences, the presence or absence of boundary allocations, and especially on the interpretation of firm shares as control rights.
3. The core is a cooperative solution notion based on blocking; the various *bargaining sets* use core logic, but impose more stringent requirements for blocking. (An allocation fails to be in the core there is a coalition C and an allocation g for C such that all members of C prefer g to f; we might say the coalition C has an *objection* to f. An allocation f fails to be in the *classical bargaining set* if there is a coalition C and an allocation g for C such that all members of C prefer g to f – so C has an objection to f – and in addition there is no coalition D and allocation h for D such that all members of D prefer h to f and all members of $C \cap D$ prefer h to g – so no coalition has a *counter-objection* to g.) The bargaining sets are larger than the core, so convergence of bargaining sets to the set of competitive

allocations is a more stringent test of competition than is convergence of the core. Anderson (1998) establishes a convergence result for the classical bargaining set and a variant, but Anderson, Trockel and Zhou (1997) show that core convergence can fail for other bargaining sets. Convergence of the core is sometimes interpreted as the inability of groups to manipulate the outcome; non-convergence of (some) bargaining sets casts doubt on this interpretation.

4. If the number of commodities is fixed and finite and the number of traders is large, then there are many potential buyers and sellers of each good; markets are *thick*. Recent work on the core in economies with a continuum of agents and an *infinite* number of commodities highlight the importance of thick markets for perfect competition. If the commodity space is separable, then there are 'many more agents than commodities', and Aumann's core equivalence theorem obtains (Rustichini and Yannelis 1991), but if the commodity space is sufficiently large then there can be 'more commodities than agents', and Aumann's theorem can fail (Tourky and Yannelis 2001). Ostroy and Zame (1994) focuses on the extent to which commodities are good substitutes, which can be interpreted as 'economic thickness'. If markets are economically thick then again Aumann's theorem obtains, but if markets are economically thin then Aumann's theorem can fail.
5. When the commodity space is infinite dimensional and the number of traders is finite, the situation is even subtler. The Debreu–Scarf core convergence theorem holds (Aliprantis et al. 1985), but most of the obvious analogues of the general core convergence theorems of Keiding, Dierker and Anderson fail (Anderson and Zame 1997). Thus there is a substantial difference between replica economies and general large finite economies; there can also be a substantial difference between large finite economies and continuum economies. Anderson and Zame (1997) identify two reasons for these differences: the first is that the integrability assumptions inherent in the continuum model, usually viewed as economically innocuous in the finite dimensional setting, impose economically serious restrictions in the infinite dimensional setting; the second is that various compactness properties that are inherent in the finite dimensional setting no longer obtain in the infinite dimensional setting. A different consequence of the latter fact is that there are well-behaved continuum economies for which the core is empty and no competitive equilibrium exists (Zame 1986).

The work discussed above seeks to give cooperative justifications for price- taking behaviour; a different literature seeks to give non-cooperative justifications. For exchange economies, Rubinstein and Wolinsky (1985) propose a search model in which agents enter the market, meet and trade at random, leave the market to consume, and are replaced. They argue that the steady-state outcome of the bargaining game may differ from the competitive outcome. Rubinstein and Wolinsky focus on a setting in which the number of potential buyers is different from the number of potential sellers, but posit a replacement process in which agents who match and leave the market are replaced with exact duplicates. As a result, agents on the short side of the market are unable to exercise their market power. Gale (1986a, b, 1987, 2000) offers different models, using different replacement processes, and shows that outcomes of the search/bargaining game (both in the steady state and not) *do* coincide with competitive outcomes.

For production economies, Allen and Hellwig (1986a, b) argue that Bertrand price competition leads to approximately competitive prices and outcomes if there is a large number of firms supplying a competitive consumption sector. Cheng (2002) argues that, if firms are risk averse and costs are uncertain, then Cournot quantity competition leads to approximately competitive prices and outcomes but Bertrand price competition may not.

Ostroy (1980, 1981) suggests a different approach to perfect competition, based more directly on the ability of individuals to influence prices or to favourably manipulate outcomes.

Favourable manipulation will certainly be impossible from any outcome for which each individual already extracts his or her marginal product. A formal definition is easiest for finite economies with transferable utility. For each sub-coalition S of the set N of all agents, let $v(S)$ be the maximal total utility obtainable by redistribution of the total endowment of S. (The assumption of transferable utility guarantees that this is economically sensible.) The marginal contribution of agent i to society is thus $v(N) - v(N/i)$. If x is the vector of utilities of a feasible allocation, then agent i extracts his/her marginal product if $x_i = v(N) - v(N/i)$; an allocation at which each agent extracts his or her marginal product is said to satisfy the *no-surplus* property. The appropriate extensions of these definitions to continuum economies use limits of small coalitions as proxies for individuals and derivatives as proxies for individual marginal products. For continuum economies with a finite number of goods, no-surplus is generic – but not universal. No-surplus is closely related to perfect elasticity of demand and supply (Ostroy 1984) and to the inability of agents to manipulate; for some environments, these notions are coincident (Gretsky et al. 1999). Makowski and Ostroy (2001) provide an overview, bibliography and applications to innovation and mechanism design; Makowski (2004) gives a striking application to non-contractible investment and the 'hold-up' problem. See also perfect competition; Shapley value.

Equilibration

The definition of competitive equilibrium identifies a particular state of the economy but provides no clue as to the process by which the economy is to reach this state. Without such a process, competitive equilibrium may retain its usefulness as a benchmark (normative solution), but is in doubt as a description of reality (positive solution). Unfortunately, no such process has been described.

Walras suggested a simple and appealing process that he called *tâtonnement:* from any given price system p, adjust prices in proportion to excess demands. However, for some economies, the tâtonnement process does not converge (Scarf 1960). Indeed, in view of the fact that, aside from the necessity of satisfying Walras's law, the excess demand function of an economy is essentially arbitrary (Sonnenschein 1973; Mantel 1974; Debreu 1974), the tâtonnement process may follow an essentially arbitrary dynamic: converge from some initial prices, diverge from others, and cycle from still others. Although more complicated adjustment processes have been proposed, none seems economically sensible. Moreover, any adjustment process that is universally convergent must of necessity use an enormous amount of information: not just the excess demand of each good at each price, but the derivatives of excess demand of each good with respect to own price *and* the prices of other goods as well (Saari and Simon 1978; Saari 1985). An additional difficulty with tâtonnement is that, because prices adjust without trade, it does not seem to describe any process we see in the real world. (Walras imagined a fictitious 'auctioneer' who sets a tentative price, receives tentative demands, adjusts the tentative price, and so on, with the process continuing until excess demand for all goods is zero, at which point trade takes place.)

Keisler (1996) offers a suggestive model of price adjustment *with* trade. Consider a large finite population of traders and a single warehouse. The warehouse manager announces an initial price p_0. At each time t thereafter, a single consumer, chosen at random, comes to the warehouse, trades at the current price p_t, and leaves the economy. (Thus, the trader has no incentive to misrepresent or to wait.) The warehouse manager adjusts prices in a direction proportional to the net trade of the most recent consumer, leading to a new price p_{t+1}, and the process continues. Keisler shows that, if the population is large, the price adjustments (the constant of proportionality) are small, and the initial price p_0 is in the basin of attraction of some price p^* that is a stable equilibrium for the Walrasian tâtonnement process, then with probability close to 1 the price path will reach, and eventually stay in, a small neighbourhood of p^* and most trades will take place at prices near p^*. (Because the warehouse

manager adjusts prices following every trade, prices do not converge – but they do not leave a small neighbourhood of p^*.)

Given that the Walrasian tâtonnement process need have no stable equilibria, one might ask why Keisler's result is of interest. It should be kept in mind, however, that the Debreu–Mantel–Sonnenschein theorem describes only the *theoretical possibilities* for the aggregate excess demand function of an economy; it does not describe the aggregate excess demand function of any real economy. If the aggregate excess demand function of the real economy is – always or frequently – well- behaved, then Keisler's result provides hope for a sensible process that converges to equilibrium.

Of course it is not possible to observe the aggregate excess demand function of the real economy. Failing that, it seems natural to ask whether there are reasonable conditions on preferences and endowments – and especially on the *distribution* of preferences and endowments – that are compatible with empirical observation and also guarantee that the aggregate excess demand function is well-behaved (stable or locally stable for Walrasian tâtonnement or some other natural adjustment process). A sufficient condition for this to be true is that the economy admit a representative consumer, in the sense that the demand function of the economy is the demand function of a one-agent economy. (The tâtonnement process follows the differential equation $dp/dt = K[w - D(p)]$, where w is the aggregate endowment and $D(p)$ is aggregate demand at the price p, given endowment w. Fix any equilibrium price p; by definition, $D(p^*) = 0$. At any non-equilibrium price p, w is affordable but not chosen, so w is dis-preferred to $D(p)$. By assumption, D is the demand function of a single agent, so revealed preference applies; hence $D(p)$ cannot be affordable at the equilibrium price p^*; that is, $p^* \cdot D(p) > 0$. Walras's Law guarantees that $p \cdot D(p) = p \cdot w$ for all prices p. Taken together, these are enough to guarantee that the tâtonnement process converges from any initial price to the equilibrium price p^*.)

One promising approach focuses on the distribution of preferences and endowments and shows that aggregate market demand D obeys the *law of demand;* that is, $(p - p') \cdot (D(p) - D(p')) < 0$. (Existence of a representative consumer is not enough to not guarantee the law of demand in the aggregate.) For example, if all agents have the same demand function, income is independent of price, the income density is decreasing and the smallest incomes are sufficiently small, then the law of demand will hold in the aggregate (Hildenbrand 1983). These assumptions are strong but can be significantly weakened (Chiappori 1985; Quah 1997). Alternatively, if all agents have the same income, then sufficient heterogeneity of demand functions will also imply the law of demand in the aggregate (Grandmont 1987); again, these assumptions are strong, but can be significantly weakened (Grandmont 1992; Quah 2002). See also aggregation (theory).

Incomplete Markets

The standard Arrow–Debreu–McKenzie general equilibrium model posits a market for every commodity. Since the description of a commodity includes the date and state of nature in which it will be delivered, this entails markets for all claims to all goods in all future dates and states of the world. Radner (1972), building on a earlier model of Arrow (1953), offers an alternative model in which at each date and state of the world there are spot markets for commodities available at that date and state of the world and for *assets* or *securities* (state-contingent claims to wealth at future dates), but not for commodities at other dates and states. In this model, the transfer of wealth across time or across states of nature can be accomplished only by trading available assets. If all state- and time-dependent wealth transfers can be accomplished by trading available assets, then asset markets are *complete,* and the model reduces to the Arrow–Debreu–McKenzie model; in the alternative case, asset markets are *incomplete.*

In principle, asset payoffs (or *dividends)* in a given date-event may depend arbitrarily on commodity spot prices in that date-event, and even on the prices of other assets. Two particular kinds of

securities are of special interest. *Financial assets* or *nominal assets* are those whose dividends are independent of prices; such assets are abstractions of real-world instruments such as treasury bills. *Real assets* are those whose dividends in a given date-event are the value at commodity spot prices of a specified bundle of commodities; such assets are abstractions of real- world instruments such as commodity forward contracts. (Most real-world forward contracts are *marked-to-market;* that is, they promise to deliver the *value* of a particular bundle of commodities rather than the physical bundle itself. In a perfectly competitive market, the distinction is unimportant, but in a real-world market the distinction can be significant if, as sometimes happens, the physical good promised is in sufficiently short supply that the promised quantity cannot be delivered.)

Radner (1972) establishes the existence of an asset market equilibrium (for either nominal or real assets), assuming an exogenously given bound on short sales. Such a constraint is unsatisfactory because if the constraint is binding then the equilibrium depends on an arbitrarily given, perhaps not economically meaningful, bound. For economies in which all assets are nominal, Cass (1984), Werner (1985) and Duffie (1987) show that short sale bounds are not necessary. The case of real assets is more subtle, because the possibilities for wealth transfer may depend on commodity spot prices. Suppose, for example, that trading takes place today and tomorrow, that there are two possible states – rainy and sunny – of the world tomorrow, and that there are only two assets, one promising delivery of (the value of) one bushel of wheat in each state, the other promising delivery of (the value of) one bushel of corn in each state. If the ratio of the price of wheat to the price of corn is different in the rainy state than in the sunny state, then the dividends of these assets are linearly independent, and the market is complete; if the ratio of the price of wheat to the price of corn is the *same* in the rainy state as in the sunny state, then these assets are collinear and the market is incomplete. In the former case, the space of wealth patterns that can be achieved by trading assets is two-dimensional; in the latter case, it is one-dimensional. In particular, consumers' budget sets are *discontinuous* functions of commodity spot prices. (For financial assets, dividends are by definition independent of prices, so this phenomenon cannot arise.) As Hart (1975) shows, this phenomenon leads to examples in which no equilibrium exists. However, Duffie and Shafer (1985, 1986) show that equilibrium does exist for *generic* values of the parameters. (The proof uses degree theory, because familiar arguments based on the Kakutani fixed point theorem are not applicable. Elegant fixed point proofs were later discovered by Husseini et al. (1990) and Hirsch et al. (1990). Geanakoplos and Shafer 1990, gave an equally elegant homotopy argument.) For more complicated assets, such as options, equilibrium may fail to exist for an open set of parameters (Ku and Polemarchakis 1990). The volume by Magill and Quinzii (1996) presents an excellent extended discussion and bibliography.

The incomplete markets model is attractive in part because it provides a framework in which to model and address many interesting economic phenomena and questions. For instance:

- When asset markets are incomplete, equilibrium commodity allocations need not be Pareto optimal. Indeed, Geanakoplos and Polemarchakis (1986) show that equilibrium commodity allocations will typically fail to be even constrained optimal: a social planner could improve welfare of all participants, even if constrained to use only existing asset markets to transfer wealth.
- Enlarging the set of available assets increases trading opportunities, so it is tempting to believe that it improves welfare. A surprising example due to Hart (1975) shows that enlarging the set of available assets may make *everyone* worse off; Elul (1995) and Cass and Citanna (1998) show that the possibility of such Pareto worsening is a robust phenomenon.
- By definition, the dividends of financial assets are independent of commodity spot prices – but the purchasing power of these dividends may depend on price levels. In a market with only financial assets, there is nothing to connect price levels in one date-event to

price levels in another, so equilibrium asset prices and purchasing power are generally indeterminate; this leads to robust indeterminacy of equilibrium prices and consumptions as well (Balasko and Cass 1989; Geanakoplos and Mas-Colell 1989). For real assets, the purchasing power of dividends is independent of price levels, and equilibrium prices and consumptions are generically determinate (Geanakoplos and Polemarchakis 1986).

- Default is suggestive of disequilibrium and inefficiency. Dubey et al. (2005) show, to the contrary, that default is compatible with equilibrium and may in fact promote welfare. Zame (1993) uses a similar framework to underscore the positive role played by default in expanding the effective span of assets when markets are incomplete.

A recurring theme in the study of asset markets is the importance of re-trading long-lived assets. The power of frequent trading underlies both the celebrated option- pricing formula of Black and Scholes (1973) and portfolio insurance (Leland 1980). (Cox et al. 1979, present a more easily understood discrete time version.) In a general discrete time model, Kreps (1982) argues that, if the number of long-lived assets is at least as great as the degree of uncertainty from one date-event to the next, then it will generically be possible to replicate any wealth pattern by frequent trading of a few long-lived assets. In particular, asset market equilibrium will, in this circumstance, coincide with complete markets equilibrium. Duffie and Huang (1985) identify an appropriate of Kreps's spanning condition in the continuous time setting (the setting most used in finance) and proves the corresponding generalization of Kreps's dynamic completeness result.

In an infinite-horizon setting, the existence of equilibrium when markets are incomplete requires ruling out the possibility of perpetual borrowing to pay each period's debts; the 'right' conditions (which may be expressed as transversality conditions, as limits on short sales, or as debt constraints) were identified (independently) by Magill and Quinzii (1994), Levine and Zame (1996) and Hernandez and Santos (1996).

The infinite-horizon setting is of particular interest because of its connection with asset-pricing. Mehra and Prescott (1985) show that historical returns on equity (stocks) cannot be reasonably explained within a complete-markets, infinite-horizon, asset-pricing model in the style of Lucas (1978). US data, (real) returns on safe assets (such as Treasury bills) are about one per cent and returns on equity are about seven per cent. With reasonable choices for time preference and risk aversion, the model suggests that returns on safe assets should be about two to three per cent and that returns on equity should be about three to four per cent; even extreme specifications of risk aversion do not yield an equity premium (that is, a rate of return on equity in excess of the return on safe assets) above two per cent. For overviews see Kocherlakota (1996) and Cochrane (2001).

A plausible objection to the complete markets assumption made by Mehra and Prescott is that labour income is not readily tradable; however, computationally tractable models with plausible parametrizations of untradable labour income do not appear to deliver a substantially higher equity premium (Telmer 1993; Lucas 1994; Heaton and Lucas 1996). On the other hand, Constantinides and Duffie (1996) show that an *arbitrary* equity premium can be generated if enough income is untradable. More precisely, given an aggregate income process and *any* system of asset prices that are consistent with time preference, there is a distribution of untradable income for which the given prices constitute an equilibrium. The argument of Constantinides and Duffie relies on individual shocks that are permanent; whether that assumption is necessary and whether the true distribution of labour income (or other untradable income) is sufficient to generate the observed equity premium is a subject of considerable interest (Levine and Zame 2002; Cogley 2002; De Santis 2005).

An interesting alternative explanation for the divergence of observed asset prices from theoretical predictions is that markets are complete but participation is not. Alvarez and Jermann (2000), building on a model of Kehoe and Levine (1993), explore the implications for asset pricing of an

environment in which imperfect enforcement generates endogenously incomplete participation. See also incomplete markets.

Infinitely Many Commodities

The description of a commodity includes its physical characteristics and the date, location and state of the world at which it will delivered. If time is modelled as continuous or the horizon is modelled as infinite, if uncertainty is modelled by the use of an infinite state space, or if commodities are modelled as having a continuous range of possible characteristics, then the number of commodities will be infinite. Examples include:

- the use (especially in macroeconomics and asset pricing) of l^∞ (the space of bounded sequences) to model of consumption and trade over an infinite time horizon (Bewley 1972; Lucas 1978; Mehra and Prescott 1985);
- the use (especially in finance) of $L^2(\Omega, F, P)$ (the space of random variables with finite mean and variance on some probability space (Ω, F, P)) (and related spaces) to model choice and asset trading under uncertainty (Black and Scholes 1973; Merton 1973; Duffie and Huang 1985);
- the use of $C[0, T]$ (the space of continuous functions on the interval $[0, T]$) or $L^\infty(\mathbb{R}_+)$ (the space of bounded functions on the non-negative real numbers) to model consumption and trade in a continuous-time framework (Gabszewicz 1968; Bewley 1972); and
- the use of $M(K)$ (the space of (signed) measures on some space K of commodity characteristics) to model finely differentiated commodities (Mas-Colell 1975; Dixit and Stiglitz 1977; Hart 1979, 1985a, b; Jones 1984).

The Arrow–Debreu–McKenzie framework is powerful because it can be applied in many different economic environments, so it is natural to look for an extension of this framework which applies in the models above (and hopefully in many others).

In looking for such an extension, several problems arise. The most obvious problem is that neither budget sets nor feasible sets need be compact (in the given topology); thus the existence of optimal choices is immediately in doubt. An approach that avoids this difficulty, and is quite generally applicable, is to make use of the fact that most infinite dimensional vector spaces admit many topologies. For example, $L^\infty = L^\infty(\mathbb{R}_+)$ admits a norm topology (where $\|f\|$ is the essential supremum of $|f|$), but it also admits two weaker topologies that arise from viewing L^∞ as the space of continuous linear functionals on $L^1 = L^1(\mathbb{R}_+)$ (the space of integrable functions on $\mathbb{R}_+$): $\sigma(L^\infty, L^1)$ (the *weak topology*) is the weakest vector space topology on L^∞ for which L^1 is the space of continuous linear functionals, and $\tau(L^\infty, L^1)$ (the *Mackey topology*) is the strongest vector space topology for which L^1 is the space of continuous linear functionals. The weak topology is weaker than the Mackey topology which is in turn weaker than the norm topology (Dunford and Schwartz 1957). The applicability to equilibrium analysis comes from three facts:

(i) in the weak topology, closed and bounded sets are compact;
(ii) convex sets that are closed in the Mackey topology are also closed in the weak topology;
(iii) for preference relations, continuity in the Mackey topology can be interpreted as impatience.

In view of (iii), Mackey continuity of preferences is an economically meaningful and natural assumption; in view of (ii), preferences that are convex and Mackey continuous are also weakly upper hemi-continuous; in view of (i), such preferences admit optimal choices whenever feasible sets are closed and bounded. (This interplay between topologies is a familiar functional-analytic theme.)

The second problem that arises is that sensible-looking preference relations may not admit supporting prices. When the commodity space is finite-dimensional and preferences are convex, the separation theorem guarantees that a

supporting price at a consumption bundle x can be constructed as a linear functional separating x from the set $p(x)$ of bundles strictly preferred to x. When the commodity space is infinite dimensional, the separation theorem only guarantees the existence of a linear functional separating x from $p(x)$ in case $p(x)$ has non-empty interior. In many spaces, this is problematical because the positive cone (the most natural consumption set) has empty interior, whence, *a fortiori*, the strictly preferred set also must have empty interior.

Mas-Colell (1986) shows by example that supporting prices need not exist, and identifies a class of preferences for which supporting prices *do* exist. Say that a preference relation defined on the positive cone X_+ of some commodity space X is *uniformly proper* if there is some vector $v \in X_+$ and some open cone C containing v such that for each $x \in X_+$ every bundle in $(x - C) \cap X_+$ is strictly dis-preferred to x. (The non-existence of supporting prices may be interpreted as unboundedness of marginal rates of substitution; uniform properness may be interpreted as a bound on marginal rates of substitution.) For commodity spaces that are topological vector lattices (ordered topological vector spaces in which every pair of elements have an infimum and a supremum and in which the lattice operations are continuous), Mas-Colell shows that uniform properness is the crucial additional assumption needed to guarantee the existence of competitive equilibrium for exchange economies with a finite number of agents. Mas-Colell (1986) and Zame (1987) extend the existence theorem to include production economies, introducing (necessary) additional conditions that bound marginal rates of transformation as well as marginal rates of substitution. Mas-Colell and Richard (1991) offer a different proof that makes weaker assumptions by focusing more on the lattice structure of the price space and less on the lattice structure of the commodity space; this is important in a number of applications. Mas-Colell and Zame (1991) and Aliprantis et al. (1989) provide good surveys of the existence theorems, including discussion of a number of different proof strategies.

A surprising aspect of the analysis in the infinite dimensional setting is that it relies heavily on the order structure of the commodity and prices spaces and on the assumption that consumption sets coincide with the positive cone of the commodity space, which play no role in the finite dimensional setting. That the order structures should play an important role is suggested by Aliprantis and Brown (1982), but a foreshadowing can already be seen in Bewley (1972). In that paper, the role of the order structure is not to guarantee the existence of equilibrium, but to guarantee the existence of equilibrium with economically meaningful prices (that is, prices in l^1 rather than in the full dual of l^∞). Bewley's argument rests on the possibility of decomposing a socially feasible bundle dominated by the social endowment into a sum of individually feasible bundles dominated by individual endowments. That this is possible is a consequence of the *Riesz decomposition property,* which holds when the commodity space is a vector lattice and consumption sets are the positive cone, but not for general commodity spaces or consumption sets. The arguments used by Mas-Colell to construct prices that support a Pareto optimal allocation, and by Yannelis and Zame (1986) to provide estimates on supporting prices, make similar use of the Riesz decomposition property and so again require that the commodity space be a lattice. For explorations of equilibrium theory when assumptions on the order structure are relaxed, see Aliprantis et al. (2001) and Aliprantis et al. (2005). See also functional analysis.

Hidden Information and Hidden Actions

The standard model of competitive markets treats an environment in which all agents are equally informed about economically relevant parameters, and has nothing to say about environments in which agents are asymmetrically informed – but even casual observation suggests that the latter environments are more common than the former.

That asymmetric information can have an enormous impact on market outcomes is pointed out forcefully in Akerlof's (1970) seminal discussion of used-car markets. Because sellers typically have private information about their own cars,

such markets display *adverse selection*: low-quality cars are offered for sale more readily than are high-quality cars, so the quality distribution of cars offered for sale at a given price will be skewed downward in comparison with the overall distribution of cars in the market. As a result, market outcomes may be less efficient than in the case where all information is public; in extreme situations, only autarkic outcomes (no trade) may obtain, even though every potential buyer values every car more than its original owner.

Adverse selection arises in many other markets as well. Potential borrowers may know more about their creditworthiness than do potential lenders, owners/ operators of a productive firm may know more about its future profitability than potential investors, and potential buyers of insurance may know more about their accident or health risks than do potential sellers of insurance. In the insurance context, Rothschild and Stiglitz (1976) argue that adverse selection may become so important that equilibrium does not exist. (But Rothschild and Stiglitz use a mixed, and not strictly price-taking, notion of equilibrium.)

The work of Akerlof makes it clear that asymmetric information may matter for market outcomes; the work of Rothschild and Stiglitz makes it clear that asymmetric information may matter for the way we model markets as well. Following these seminal contributions, a large literature has sought to integrate asymmetric information with general equilibrium modelling of competitive markets. Central issues in this literature include: which models are appropriate for which kinds of asymmetric information? Does the operation of the market reveal information? If so, how? And how much?

Radner (1968) develops a model of an environment in which agents learn their private information (modelled as an information partition over true states of the world) and make contingent trades *before* the true state of the world occurs. In this model, private information acts as a constraint on choices: each agent can choose only among state-contingent consumption bundles that are measurable with respect to his/her private information; call such bundles *private.* A *Walrasian expectations equilibrium* consists of economy-wide prices and private consumption bundles for each agent such that each agent's bundle is optimal among private, budget-feasible consumptions, and markets clear in each state of the world. If free disposal is permitted, standard conditions guarantee that Walrasian expectations equilibrium exists. For these environments, Walrasian expectations equilibrium can be justified as a descriptive theory in much the same way that Walrasian equilibrium is justified for symmetric information environments: as the limit of a cooperative solution. For these asymmetric information environments, the appropriate cooperative notion, the *private core*, consists of private allocations (vectors of private consumption choices) which have the property that no coalition can construct an improving private allocation, using only its own resources (Yannelis 1991; Allen 1994, 2003. Aliprantis et al. (2001), Einy et al. (2003) and Hervés-Beloso et al. (2005) show that versions of Debreu and Scarf's (1963) core convergence theorem and Aumann's (1964) core equivalence theorem obtain for the private core and Walrasian expectations equilibrium.

For environments in which trade takes place *after* the true state of nature occurs but before it is publicly known, agents may use their own private information *and* draw inferences from market activities, such as prices (Radner 1979). A *rational expectations equilibrium* consists of a (state-dependent) price function and an allocation such that each agent's state-dependent consumption bundle is measurable with respect to the join of his or her own information and the information in prices, each agent optimizes among measurable budget-feasible bundles, and markets clear in each state of the world. A simple example (Kreps 1979) shows that some simple economies do not admit any rational expectations equilibrium. However, for generic specifications of economic parameters (endowments, information and preferences or utility functions) there is always a rational expectations equilibrium in which all information is fully revealed (Allen 1981). In the presence of noise, such as uncertainty about aggregate endowments, equilibrium may be partially, but not fully, revealing (Admati 1991). Rational expectations

equilibrium forms the theoretical basis of a cornerstone of finance, the *efficient markets hypothesis,* which asserts that all information is revealed in prices.

An alternative view of rational expectations equilibrium, as a rest point of a rational, but imperfect, learning process, is offered by Anderson and Sonnenschein (1985). In their framework, each agent has an exogenously given (linear) model of the world and the economy, and chooses parameters of the model to best fit the observed data. At equilibrium (a rest point of the process of fitting parameters to data), each agent's model is best fitting but not necessarily correct. At such an equilibrium, agents may not have learned everything, but they have learned all that is possible for them to learn, given their models. Bossaerts (2002) addresses the econometric implications of this kind of rational, but imperfect, learning process.

A criticism of rational expectations equilibrium is that it does not address the mechanism through which agents obtain their private information. This seems an important omission because, if all information were to be revealed by prices, there would seem to be no incentive for agents to acquire information in the first place, especially if acquiring information is costly (Grossman and Stiglitz 1980). A second criticism of rational expectations equilibrium is that extracting information from prices seems to require agents to have a great deal of information about the economy (including information about other agents); when equilibrium is fully revealing, for example, agents must be able to invert the map from states of the world to equilibrium prices. Perhaps the most serious criticism of rational expectations equilibrium is that it provides no process by which information gets into prices. If agents use information in prices in forming their demands, how do those demands influence prices? If demands do not influence prices, where do prices come from?

A very fruitful approach to the revelation of private information takes as its starting point the observation that private information gives rise to incentive problems: if private information is valuable, agents may not wish to take actions that reveal that private information. However, incentive problems may not matter much if agents are *informationally small.* (A new-car variation of Akerlof's familiar used-car market provides an intuitive idea of what it means to be informationally small. Suppose that all cars have the *same* true quality, unknown to both buyers and sellers, but that sellers receive noisy signals of this common quality. If there are many sellers, and sellers' signals are conditionally independent, then the marginal amount of information revealed by the signal of a given seller is small. Put differently: in an economy with many sellers, each seller's signal has little effect on the true posterior information about quality.) This idea can be formalized in a number of different ways. For instance, Gul and Postlewaite (1993) and McLean and Postlewaite (2002) describe classes of environments for which allocations close to competitive allocations are incentive compatible in large replications. McLean and Postlewaite (2005) define an *incentive-compatible core* and show that incentive compatible core allocations in replica economies are close to competitive allocations of the full information economy provided that agents are informationally small. Forges, Heifetz and Minelli (2001) prove a related convergence result in a different formulation that includes lotteries.

Prescott and Townsend (1984a, b) pioneered an approach that treats certain kinds of private information and incentive issues within the standard general equilibrium paradigm. Prescott and Townsend model shocks (hidden information) as affecting preferences. Incentive problems (moral hazard) arise in guaranteeing that agents who experience a particular shock should have no incentive to misrepresent themselves as having experienced a different shock; these incentive constraints are incorporated into consumption sets. Two unusual aspects of the model is that objects of choice are *lotteries* over consumption bundles and that prices are linear in probabilities but not necessarily in consumption. Competitive equilibria exist and are Pareto optimal within the class of allocations that satisfy the incentive constraints. (Allowing for lotteries guarantees that consumption sets and preferences are convex, so that familiar Arrow–Debreu existence results can

be applied. It does not seem commonly observed that, in the continuum of agents framework Prescott and Townsend adopt, convexity of consumption sets and preferences is not necessary to guarantee the existence of an equilibrium.)

If enforcement of contractual arrangements is imperfect, then issues of moral hazard and adverse selection arise in financial markets as well. Borrowers may choose to default on promises (moral hazard) and borrowers who are poor risks or less affected by sanctions are more likely to default than are borrowers who are good credit risks (adverse selection). Surprisingly, neither of these interferes with the existence equilibrium, provided that deliveries on asset promises are *pooled* (Dubey et al. 2005). The most familiar pooled financial instruments are collateralized mortgage obligations, which are pools of individual mortgages. Pooling mortgage deliveries spreads the default risk across all lenders; absent pooling, each lender would face the idiosyncratic risk that individual borrowers default against them. Although default is suggestive of deadweight losses, allowing for default may be Pareto-improving when markets are incomplete, because it expands the effective span of available assets (Dubey et al. 2005; Zame 1993).

Bisin et al. (2002) argue that, if *all trade* – even trade for commodities – is carried out through contracts, and deliveries on all contracts are pooled, then both adverse selection and moral hazard can be accommodated within almost standard general equilibrium models. (The assumption that deliveries on commodity contracts are pooled would seem natural for orange juice, but not for used cars; the assumption that deliveries on financial contracts are pooled would seem natural for collateralized mortgage obligations but not for individual mortgages.) Dubey and Geanakoplos (2002) use a similar idea to reformulate the insurance economy of Rothschild and Stiglitz (1976) and argue that an equilibrium always exists.

All of the work described above considers either economies in which individuals consume by themselves and production (if any) is of the standard Arrow–Debreu–McKenzie type. A growing literature treats economies in which individuals consume and/or produce in small groups (teams or firms). Prescott and Townsend (2006), Rahman (2005) and Song (2006) treat general equilibrium models with team production. Output is a function of observable investment and unobservable effort, which creates a moral hazard problem. Working in the tradition of Prescott and Townsend, these papers find institutional arrangements that permit the decentralization of incentive-efficient configurations. When production is deterministic and utility is transferable, the requisite institutions include contract arbitrageurs and Lindahl (personalized) prices for team membership; when production is stochastic and utility is not transferable, public randomization devices and assets whose payoffs depend on the distribution of idiosyncratic uncertainty are required as well.

Zame (2005) takes a different approach, adding hidden information and hidden actions to the clubs framework of Ellickson et al. (1999, 2006). In that model, firm output and individual utility depend on skills and actions of other agents, which are unobservable and uncontractable; thus there is scope for both adverse selection and moral hazard. Moreover, because output within firms depends on action profiles, agents are subject to idiosyncratic risk. The set of firms that form and the contractual arrangements that appear, the assignments of agents to firms, the prices faced by firms for inputs and outputs, and the incentives to agents are all determined endogenously at equilibrium. Agents choose consumption – but they also choose which firms to join, which roles to occupy in those firms, and which actions to take in those roles. Agents interact anonymously with the (large) market, but strategically within the (small) firms they join. The model accommodates moral hazard, adverse selection, signalling and insurance. Equilibrium allocations may be incentive efficient and even Pareto ranked.

See Also

- ▶ Adverse Selection
- ▶ Aggregation (Theory)
- ▶ Efficient Markets Hypothesis
- ▶ Functional Analysis

- ▶ Incomplete Markets
- ▶ Perfect Competition
- ▶ Shapley Value

Bibliography

Admati, A. 1991. A noisy rational expectations equilibrium for multi-asset securities markets. *Econometrica* 53: 629–658.

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

Aliprantis, C.D., and D.J. Brown. 1982. Equilibria in markets with a Riesz space of commodities. *Journal of Mathematical Economics* 11: 189–207.

Aliprantis, C.D., D.J. Brown, and O. Burkinshaw. 1985. Edgeworth equilibria. *Econometrica* 55: 1109–1138.

Aliprantis, C.D., D.J. Brown, and O. Burkinshaw. 1989. *Existence and optimality of competitive equilibria*. New York: Springer.

Aliprantis, C.D., R. Tourky, and N.C. Yannelis. 2001. A theory of value with nonlinear prices: Equilibrium analysis beyond vector lattices. *Journal of Economic Theory* 100: 22–72.

Aliprantis, C.D., M. Florenzano, and R. Tourky. 2005. *General equilibrium analysis in ordered topological vector spaces*, Working paper. CERSEM.

Allen, B. 1981. Generic existence of completely revealing equilibria for economies with uncertainty when prices convey information. *Econometrica* 49: 1173–1199.

Allen, B. 1994. Market games with asymmetric information: The core with finitely many states of the world. In *Models and experiments in risk and rationality*, ed. B. Munier and M.J. Machina. Dordrecht: Kluwer.

Allen, B. 2003. Incentives in market games with asymmetric information: The core. *Economic Theory* 21: 527–544.

Allen, B., and M. Hellwig. 1986a. Price-setting firms and the oligopolistic foundations of perfect competition. *American Economic Review* 76: 387–392.

Allen, B., and M. Hellwig. 1986b. Bertrand–Edgeworth oligopoly in large markets. *Review of Economic Studies* 53: 175–204.

Allen, B., and N. Yannelis. 2001. Differential information economies: Introduction. *Economic Theory* 18: 263–273.

Alvarez, F., and U.J. Jermann. 2000. Efficiency, equilibrium, and asset pricing with risk of default. *Econometrica* 68: 775–797.

Anderson, R.M. 1978. An elementary core equivalence theorem. *Econometrica* 46: 1483–1487.

Anderson, R.M. 1993. The core in perfectly competitive economies. In *Handbook of game theory with economic applications*, ed. R. Aumann and S. Hart, vol. I. Amsterdam: North-Holland.

Anderson, R.M. 1998. Convergence of the Aumann–Davis–Maschler and Geanakoplos bargaining sets. *Economic Theory* 11: 1–37.

Anderson, R.M., and H. Sonnenschein. 1985. Rational expectations equilibrium with econometric models. *Review of Economic Studies* 52: 359–369.

Anderson, R.M., and W.R. Zame. 1997. Edgeworth's conjecture with infinitely many commodities: L^1. *Econometrica* 65: 225–274.

Anderson, R.M., and W.R. Zame. 1998. Edgeworth's conjecture with infinitely many commodities: Differentiated commodities. *Economic Theory* 11: 331–377.

Anderson, R.M., and W.R. Zame. 2001. Genericity with infinitely many parameters. *Advances in Theoretical Economics* 1: 1–62.

Anderson, R.M., W. Trockel, and L. Zhou. 1997. Nonconvergence of the Mas-Colell and Zhou bargaining sets. *Econometrica* 65: 1227–1239.

Araujo, A. 1987. The non-existence of smooth demand in general Banach spaces. *Journal of Mathematical Economics* 17: 1–11.

Arrow, K. 1953. Le rôle des valeurs boursières pour la répartition la meilleure des risques. *Econométrie, Colloques Internationaux du C.N.R.S.* 40: 41–47.

Aumann, R.J. 1964. Markets with a continuum of traders. *Econometrica* 32: 39–50.

Aumann, R.J. 1966. Equilibrium in markets with a continuum of traders. *Econometrica* 34: 1–17.

Balasko, Y., and D. Cass. 1989. The structure of financial equilibrium with exogenous yields: The case of incomplete markets. *Econometrica* 57: 135–162.

Bennardo, A., and P.A. Chiappori. 2003. Bertrand and Walras equilibria under moral hazard. *Journal of Political Economy* 111: 785–817.

Bewley, T. 1972. Existence of equilibria in economies with infinitely many commodities. *Journal of Economic Theory* 43: 514–540.

Bisin, A., J. Geanakoplos, P. Gottardi, E. Minelli, et al. 2002. *Markets and contracts*, Discussion paper. Cowles Foundation.

Black, F., and M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 837–854.

Blume, L., and W.R. Zame. 1993. The algebraic geometry of competitive equilibrium. In *Essays in general equilibrium and international trade: In memoriam Trout Rader*, ed. W. Neuefeind. New York: Springer.

Bossaerts, P. 2002. *The paradox of asset pricing*. Princeton: Princeton University Press.

Cass, D. 1984. *Competitive equilibria in incomplete financial markets*, Working paper No. 84–09, CARES-S. University of Pennsylvania.

Cass, D., and A. Citanna. 1998. Pareto improving financial innovation in incomplete markets. *Economic Theory* 11: 467–494.

Cheng, H.C. 1996. Values of perfectly competitive economies. In *Handbook of game theory with economic applications*, ed. R.J. Aumann and S. Hart. New York: Elsevier North-Holland.

Cheng, H. 2002. Bertrand vs. Cournot equilibrium with risk averse firms and cost uncertainty. *Economic Theory* 20: 555–577.

Chiappori, P.-A. 1985. Distribution of income and the 'law of demand'. *Econometrica* 53: 109–128.

Cochrane, J. 2001. *Asset pricing*. Princeton: Princeton University Press.

Cogley, T. 2002. Idiosyncratic risk and the equity premium: Evidence from the consumer expenditure survey. *Journal of Monetary Economics* 49: 309–334.

Cole, H., and E.C. Prescott. 1997. Valuation equilibrium with clubs. *Journal of Economic Theory* 74: 19–39.

Cole, H., G. Mailath, and A. Postlewaite. 2001. Efficient non-contractible investments in large economies. *Journal of Economic Theory* 101: 333–373.

Constantinides, G., and J.D. Duffie. 1996. Asset pricing with heterogeneous traders. *Journal of Political Economy* 104: 219–240.

Cox, J.C., S.A. Ross, and M. Rubinstein. 1979. Option pricing: A simplified approach. *Journal of Financial Economics* 7: 229–263.

De Santis, M. 2005. *Interpreting aggregate stock market behavior: How far can the standard model go?* Working paper. Dartmouth College.

Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38: 387–392.

Debreu, G. 1972. Smooth preferences. *Econometrica* 40: 603–615.

Debreu, G. 1974. Excess demand functions. *Journal of Mathematical Economics* 1: 15–23.

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

Dierker, E. 1975. Equilibria and core of large economies. *Journal of Mathematical Economics* 2: 155–169.

Dixit, A., and J. Stiglitz. 1977. Monopolistic competition and optimum product diversity. *American Economic Review* 67: 297–308.

Dubey, P., and J. Geanakoplos. 2002. Competitive pooling: Rothschild-Stiglitz reconsidered. *Quarterly Journal of Economics* 117: 1529–1570.

Dubey, P., J. Geanakoplos, and M. Shubik. 2005. Default and punishment in general equilibrium. *Econometrica* 73: 1–38.

Duffie, J.D. 1987. Stochastic equilibria with incomplete financial markets. *Journal of Economic Theory* 41, 405–16. Corrigendum, 1989. *Journal of Economic Theory* 49: 384.

Duffie, J.D., and C.-F. Huang. 1985. Implementing Arrow–Debreu equilibria by continuous trading of few long-lived securities. *Econometrica* 53: 1337–1356.

Duffie, J.D., and W. Shafer. 1985. Equilibrium in incomplete markets I: A basic model of generic existence. *Journal of Mathematical Economics* 14: 285–300.

Duffie, J.D., and W. Shafer. 1986. Equilibrium in incomplete markets II: Generic existence in stochastic economies. *Journal of Mathematical Economics* 15: 199–216.

Dunford, N., and J.T. Schwartz. 1957. *Linear operators*. New York: Interscience.

Einy, E., D. Moreno, and B. Shitovitz. 2003. Competitive and core allocations in large economies with differential information. *Economic Theory* 18: 321–332.

Ellickson, B., B. Grodal, S. Scotchmer, and W.R. Zame. 1999. Clubs and the market. *Econometrica* 67: 1185–1217.

Ellickson, B., B. Grodal, S. Scotchmer, and W.R. Zame. 2006. The organization of production, consumption and learning. In *Institutions, equilibria and efficiency: Essays in Honor of Birgit Grodal*, ed. C. Schultz and K. Vind. Berlin: Springer.

Elul, R. 1995. Welfare effects of financial innovation in incomplete markets economies with several consumption goods. *Journal of Economic Theory* 65: 43–78.

Forges, F., A. Heifetz, and E. Minelli. 2001. Incentive compatible core and competitive equilibria in differential information economies. *Economic Theory* 18: 349–365.

Gabszewicz, J. 1968. Coeurs et allocations concurrentielles dans les économies d'exchange avec un continu de bains. Librarie Universitaire, Université Catholique de Louvain.

Gale, D. 1986a. Bargaining and competition. Part I: characterization. *Econometrica* 54: 785–806.

Gale, D. 1986b. Bargaining and competition. Part II: existence. *Econometrica* 54: 807–818.

Gale, D. 1987. Limit theorems for markets with sequential bargaining. *Journal of Economic Theory* 43: 20–54.

Gale, D. 2000. *Strategic foundations of general equilibrium: Dynamic matching and bargaining games*. Cambridge: Cambridge University Press.

Geanakoplos, J., and A. Mas-Colell. 1989. Real indeterminacy with financial assets. *Journal of Economic Theory* 47: 22–38.

Geanakoplos, J., and H. Polemarchakis. 1986. Existence, regularity and constrained suboptimality of competitive allocations when the asset market is incomplete. In *Uncertainty, information and communication: Essays in honor of K.J. Arrow*, ed. W. Heller, R. Starr, and D. Starrett, vol. 3. Cambridge: Cambridge University Press.

Geanakoplos, J., and W. Shafer. 1990. Solving systems of simultaneous equations in economics. *Journal of Mathematical Economics* 19: 69–93.

Geanakoplos, J., and W.R. Zame. 2002. *Collateral and the enforcement of intertemporal contracts*, Working paper. UCLA.

Ghosal, S., and H.M. Polemarchakis. 1997. Nash–Walras equilibria. *Ricerche Economiche* 51: 31–40.

Grandmont, J.-M. 1987. Distribution of preferences and the 'law of demand'. *Econometrica* 55: 155–161.

Grandmont, J.-M. 1992. Transformations of the commodity space, behavioral heterogeneity, and the aggregation problem. *Journal of Economic Theory* 57: 1–35.

Gretsky, N., J.M. Ostroy, and W.R. Zame. 1999. Perfect competition in the continuous assignment model. *Journal of Economic Theory* 88: 60–118.

Grossman, S., and J. Stiglitz. 1980. On the impossibility of informationally efficient markets. *American Economic Review* 70: 393–408.

Gul, F., and A. Postlewaite. 1993. Asymptotic efficiency in large exchange economies with asymmetric information. *Econometrica* 60: 1273–1292.

Hara, C. 2005. Existence of equilibria in economies with bads. *Econometrica* 73: 647–658.

Hart, O. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11: 418–443.

Hart, O. 1979. Monopolistic competition in a large economy with differentiated commodities. *Review of Economic Studies* 46: 1–30.

Hart, O. 1985a. Monopolistic competition in the spirit of Chamberlin: A general model. *Review of Economic Studies* 52: 529–546.

Hart, O. 1985b. Monopolistic competition in the spirit of Chamberlin: Special results. *Economic Journal* 95: 889–908.

Hart, S., and A. Mas-Colell. 1996. Bargaining and value. *Econometrica* 64: 357–380.

Heaton, J., and D.J. Lucas. 1996. Evaluating the effects of incomplete markets on risk sharing and asset pricing. *Journal of Political Economy* 104: 433–467.

Hernandez, D.A., and M.S. Santos. 1986. Competitive equilibria for infinite-horizon economies with incomplete markets. *Journal of Economic Theory* 71: 102–130.

Hervés-Beloso, C., E. Moreno-Garcia, and N.C. Yannelis. 2005. An equivalence theorem for a differential information economy. *Journal of Mathematical Economics* 41: 844–856.

Hildenbrand, W. 1983. On the 'law of demand'. *Econometrica* 51: 997–1020.

Hirsch, M., M. Magill, and A. Mas-Colell. 1990. A geometric approach to a class of equilibrium existence problems. *Journal of Mathematical Economics* 19: 95–106.

Husseini, S.Y., J.M. Lasry, and M. Magill. 1990. Existence of equilibrium with incomplete markets. *Journal of Mathematical Economics* 19: 39–67.

Jones, L. 1984. A competitive model of commodity differentiation. *Econometrica* 52: 507–530.

Kehoe, T.J., and D.K. Levine. 1985. Comparative statics and perfect foresight in infinite horizon economies. *Econometrica* 53: 433–452.

Kehoe, T.J., and D.K. Levine. 1993. Debt-constrained asset markets. *Review of Economic Studies* 60: 885–888.

Kehoe, T.J., D.K. Levine, A. Mas-Colell, and W.R. Zame. 1989. Determinacy of equilibrium in large-square economies. *Journal of Mathematical Economics* 18: 231–263.

Keiding, H. 1974. *A limit theorem on the core of large but finite economies*, Working paper. University of Copenhagen.

Keisler, J. 1996. Getting to a competitive equilibrium. *Econometrica* 64: 29–49.

Kocherlakota, N. 1996. The equity premium: It's still a puzzle. *Journal of Economic Literature* 34: 42–71.

Kreps, D. 1979. *Three essays on capital markets*. Technical report No. 298. Institute for Mathematical Studies in the Social Sciences, Stanford University.

Kreps, D. 1982. Multiperiod securities and the efficient allocation of risk: A comment on the Black–Scholes option pricing model. In *The economics of uncertainty and information*, ed. J. McCall. Chicago: University of Chicago Press.

Ku, B.-I., and H. Polemarchakis. 1990. Options and equilibrium. *Journal of Mathematical Economics* 19: 107–112.

Leland, H. 1980. Who should buy portfolio insurance? *Journal of Finance* 35: 581–594.

Levine, D.K., and W.R. Zame. 1996. Debt constraints and equilibrium in infinite horizon economies with incomplete markets. *Journal of Mathematical Economics* 26: 103–131.

Levine, D.K., and W.R. Zame. 2002. Does market incompleteness matter? *Econometrica* 70: 1805–1839.

Lucas, R.E. 1978. Asset pricing in a pure exchange economy. *Econometrica* 46: 1429–1445.

Lucas, D.J. 1994. Asset pricing with undiversifiable income risk and short sales constraints: Deepening the equity premium puzzle. *Journal of Monetary Economics* 34: 325–341.

Magill, M.J.P., and M. Quinzii. 1994. Infinite horizon incomplete markets. *Econometrica* 62: 853–880.

Magill, M.J.P., and M. Quinzii. 1996. *Theory of incomplete markets*. Vol. 1. Cambridge: MIT Press.

Magill, M.J.P., and M. Quinzii. 1997. Which improves welfare more: A nominal or an indexed bond? *Economic Theory* 10: 1–37.

Makowski, L. 2004. *Pre-contractual investment with the fear of holdups: The perfect competition connection*, Working paper. UC Davis.

Makowski, L., and J.M. Ostroy. 2001. Perfect competition and the creativity of the market. *Journal of Economic Literature* 39: 479–535.

Manelli, A. 1991a. Monotonic preferences and core equivalence. *Econometrica* 59: 123–138.

Manelli, A. 1991b. Core convergence without monotonic preferences or free disposal. *Journal of Economic Theory* 55: 400–415.

Mantel, R. 1974. On the characterization of aggregate excess demand functions. *Journal of Economic Theory* 7: 348–353.

Mas-Colell, A. 1975. A model of equilibrium with differentiated commodities. *Journal of Mathematical Economics* 2: 263–295.

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

Mas-Colell, A. 1986a. The price equilibrium existence theorem in topological vector lattices. *Econometrica* 54: 1039–1054.

Mas-Colell, A. 1986b. Valuation equilibrium and Pareto optimum revisited. In *Contributions to mathematical*

economics, ed. W. Hildenbrand and A. Mas-Colell. New York: North-Holland.
Mas-Colell, A., and S.F. Richard. 1991. A new approach to the existence of equilibria in vector lattices. *Journal of Economic Theory* 53: 1–11.
Mas-Colell, A., and W.R. Zame. 1991. Equilibrium theory in infinite-dimensional spaces. In *Handbook of mathematical economics*, ed. W. Hildenbrand and H. Sonnenschein, vol. IV. New York: Elsevier.
McLean, R., and A. Postlewaite. 2002. Informational size and incentive compatibility. *Econometrica* 70: 2421–2454.
McLean, R., and A. Postlewaite. 2005. Core convergence with asymmetric information. *Games and Economic Behavior* 50: 58–78.
Mehra, R., and E.C. Prescott. 1985. The equity premium: A puzzle. *Journal of Monetary Economics* 14: 145–161.
Merton, R.C. 1973. An intertemporal capital asset pricing model. *Econometrica* 41: 867–887.
Minehart, D.F. 1997. Generic finiteness of the set of equilibria in a finite exchange economy. *Mathematical Social Sciences* 34: 75–80.
Minelli, M., and H.M. Polemarchakis. 2000. Nash–Walras equilibria of a large economy. *Proceedings of the National Academy of Sciences (USA)* 97: 5675–5678.
Ostroy, J.M. 1980. The no-surplus condition as a characterization of perfectly competitive equilibrium. *Journal of Economic Theory* 22: 183–207.
Ostroy, J.M. 1981. Differentiability as convergence to perfectly competitive equilibrium. *Journal of Mathematical Economics* 8: 59–73.
Ostroy, J.M. 1984. A reformulation of the marginal productivity theory of distribution. *Econometrica* 52: 599–630.
Ostroy, J.M., and W.R. Zame. 1994. Non-atomic economies and the boundaries of perfect competition. *Econometrica* 62: 593–633.
Prescott, E.C., and R.M. Townsend. 1984a. Pareto optima and competitive equilibria with adverse selection and moral hazard. *Econometrica* 52: 21–45.
Prescott, E.C., and R.M. Townsend. 1984b. General competitive analysis in an economy with private information. *International Economic Review* 25: 1–20.
Prescott, E.S., and R.M. Townsend. 2006. Clubs as firms in Walrasian economies with private information. *Journal of Political Economy* 114: 644–671.
Quah, J. 1997. The law of demand when income is price dependent. *Econometrica* 65: 1421–1442.
Quah, J. 2002. The monotonicity of individual and market demand. *Econometrica* 68: 911–930.
Radner, R. 1968. Competitive equilibrium under uncertainty. *Econometrica* 31: 31–58.
Radner, R. 1972. Existence of equilibrium of plans, prices and price expectations in a sequence of markets. *Econometrica* 40: 289–303.
Radner, R. 1979. Rational expectations equilibrium: Generic existence and the information revealed by prices. *Econometrica* 47: 655–678.
Rahman, D. 2005. *Contractual pricing with incentive constraints*, Working paper. UCLA.
Rothschild, M., and J. Stiglitz. 1976. Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *Quarterly Journal of Economics* 90: 630–649.
Rubinstein, A., and A. Wolinsky. 1985. Equilibrium in a market with sequential bargaining. *Econometrica* 53: 1133–1150.
Rustichini, A., and P. Siconolfi. 2002. *General equilibrium in economies with adverse selection*, Working paper. University of Minnesota.
Rustichini, A., and N. Yannelis. 1991. Edgeworth's conjecture in economies with a continuum of agents and commodities. *Journal of Mathematical Economics* 20: 307–326.
Saari, D. 1985. Iterative price mechanisms. *Econometrica* 53: 1117–1132.
Saari, D., and C. Simon. 1978. Effective price mechanisms. *Econometrica* 46: 1097–1125.
Scarf, H. 1960. Some examples of global instability of the competitive equilibrium. *International Economic Review* 1: 157–171.
Shannon, C. 1994. Regular nonsmooth equations. *Journal of Mathematical Economics* 23: 147–166.
Shannon, C. 1999. Determinacy of competitive equilibria in economies with many commodities. *Economic Theory* 14: 29–87.
Shannon, C., and W.R. Zame. 2002. Quadratic concavity and determinacy of equilibrium. *Econometrica* 70: 631–662.
Shapley, L. 1969. Utility comparison and the theory of games. In *La Décision*. Paris: éditions du CNRS.
Song, J. 2006. *Contractual matching: Limits of decentralization*, Working paper. UCLA.
Sonnenschein, H. 1973. Do Walras' identity and continuity characterize the class of community excess demand functions? *Journal of Economic Theory* 6: 343–354.
Telmer, C.I. 1993. Asset-pricing puzzles and incomplete markets. *Journal of Finance* 48: 1803–1833.
Tourky, R., and N.C. Yannelis. 2001. Markets with many more agents than commodities: Aumann's 'hidden' assumption. *Journal of Economic Theory* 101: 189–221.
Werner, J. 1985. Equilibrium in economies with incomplete financial markets. *Journal of Economic Theory* 36: 110–119.
Xiong, S., and C. Zheng. 2005. *Core equivalence theorem with production*, Working paper. Northwestern University.
Yannelis, N.C. 1991. The core of an economy with differential information. *Economic Theory* 1: 183–198.
Yannelis, N.C., and W.R. Zame. 1986. Equilibria in Banach lattices without ordered preferences. *Journal of Mathematical Economics* 15: 85–110.
Zame, W.R. 1986. *Economies with a continuum of traders and infinitely many commodities*, Working paper. Buffalo: SUNY.

Zame, W.R. 1987. Competitive equilibria in production economies with an infinite dimensional commodity space. *Econometrica* 55: 1075–1108.
Zame, W.R. 1993. Efficiency and the role of default when security markets are incomplete. *American Economic Review* 83: 1142–1164.
Zame, W.R. 2005. *Incentives, contracts and markets: a general equilibrium theory of firms*, Working paper. UCLA.

General Equilibrium with Incomplete Markets

Michael Magill and Martine Quinzii

Abstract

An account is given of the principal concepts and results of general equilibrium with incomplete financial markets over a finite horizon, focusing on the generic existence, suboptimality and determinacy of equilibrium. Many results depend on the nature of the financial securities, whether they are real or nominal, nominal securities leading to the analysis of a class of monetary equilibrium models.

Keywords

Arrow, K.; Determinacy of equilibrium; General equilibrium; Incomplete financial markets; Spot-financial market equilibrium

JEL Classifications

D52; D53; E44; E52; G12

One of Adam Smith (1776)'s beautiful insights is that markets coordinate the activities of agents and lead to optimal allocations, even though agents act purely in their own self-interest. The idea was formalized in elegant form some 200 years later in the 1950s by Arrow, Debreu and McKenzie in the branch of economics which came to be known as *general equilibrium theory* (GE). The GE model, which involved a finite number of consumers, firms and goods, was static. Arrow (1953) and Debreu (1959) showed how the model could be extended to a setting with time and uncertainty by introducing an event-tree to describe the uncertainty, and a structure of markets in which contingent contracts for future delivery of commodities are traded at an initial date. Although this model, which has come to be known as the *Arrow-Debreu model* (AD), involves time and uncertainty in the characteristics of the economy, it is still essentially static: all trading is assumed to take place at an initial date, and at subsequent dates, promises are delivered but no new contractual commitments are made.

Spot-Financial Market Equilibrium

In a striking paper Arrow (1953) showed that any AD equilibrium could be achieved by using an alternative and more realistic sequential system of markets, consisting of financial (Arrow security) markets and spot markets for goods at each date-event. An *Arrow security* purchased or sold at date t is a contract promising to deliver one unit of income in one of the possible contingencies that can occur at date $t + 1$. If at each date-event there exists a complete set of such contracts, one for each contingency that can occur at the following date, then an AD equilibrium allocation can be achieved by a combination of these Arrow security markets for redistributing income, and spot markets for exchanging goods. When the Arrow securities are replaced by a general class of financial securities calling for the delivery of income or goods at future date-events, we obtain the concept of a *spot-financial market equilibrium*. In order that the allocation obtained with this structure of markets coincide with the allocation obtained with Arrow–Debreu contingent markets, two conditions must be satisfied: the financial markets must be *complete*, and agents must *correctly anticipate* at the initial date the spot prices of every good and the payoff of every security at every date-event in the future. This correct-anticipation condition is needed in order that the income that agents choose to bring forward by their holding of financial securities permits them to buy the bundle of goods that they had planned

to consume when choosing their income transfers. To obtain such a well-coordinated outcome agents should have familiarity with the functioning of the markets, and some stationarity in the structure of the economy should prevail in order that agents can form such correct anticipations.

Removing the assumption of correct anticipations leads to the theory of *temporary equilibrium,* which focuses on the minimal conditions on agents' expectations of future prices which permit current markets to clear. Maintaining the assumption of correct anticipations of future prices while dropping the assumption that financial markets are complete leads to the theory of general equilibrium with incomplete markets, GEI for short. The GEI model has served to improve our understanding of the relationship between the real, financial and monetary sectors of the economy by providing a common framework for studying traditional price theory, the theory of finance and monetary theory.

One of the significant contributions of general equilibrium to economic theory is that it has revealed the deep insights that an abstract and rigorous mathematical model can provide into the functioning of an economic system: rigour, abstraction and clarity of thought are the hallmarks of the GE approach. Three properties of an equilibrium – existence, optimality and determinacy – have provided the basic template for organizing the theory. Establishing existence ensures that the different components of the model fit together in a coherent way; the analysis of optimality evaluates the efficiency of the underlying market structure as a mechanism for allocating resources; establishing determinacy provides a measure of the ability of the model to predict the outcome of equilibrium. Following this programme in the theory of incomplete markets has required new mathematical techniques to analyse the properties of equilibrium. For, unlike in traditional GE, many properties of a GEI equilibrium are 'almost always true' but admit some counterexamples. For example, if the financial markets are incomplete, for almost all economies risk sharing will not be optimal: however there are some special economies studied in finance, like the mean-variance economies of the capital asset pricing model, in which the equilibrium is optimal with only bond and equity contracts which technically do not constitute a complete security structure. As a result the analysis of the GEI model relies heavily on the use of differential topology, which is the branch of mathematics ideally suited to study typical, or 'generic', properties of solutions to a system of equations.

The GEI Model

To set the stage for studying the properties of the GEI model, consider the simplest version of the model, a two-period ($t = 0,1$) exchange economy with L commodities and I agents, where each agent is uncertain about his endowment of the goods at date 1. Let 'uncertainty' be expressed by assuming that 'nature' will draw one of S possible 'states of nature', say $s \in \{1, \ldots, S\}$, and though each agent does not know which state will be chosen, he does know what his endowment $\omega_s^i = \left(\omega_{s1}^i, \ldots, \omega_{sL}^i\right)$ will be if state s occurs. For convenience we label date 0 as state 0; then agent i's endowment is $\omega^i = \left(\omega_0^i, \omega_1^i, \ldots, \omega_S^i\right), i = 1, \ldots, I$. Agents can exchange goods and share their risks by trading on spot markets (one for each good in each state), and can redistribute their income over time and across states (thereby sharing risks) by trading on financial markets.

Let p_{sl} denote the spot price of good l in state s, and let $p_s = (p_{s1}, \ldots, p_{sL})$ denote the vector of spot prices in state s; then $p = (p_0, \ldots, p_s)$ denotes the vector of spot prices across all date-events in this two-period setting. A similar notation is used for allocations.

At date 0 there are also J securities ($j = 1, \ldots, J$) that agents can trade. Security j is a promise made at date 0 to pay V_s^j if state s occurs, where the payment V_s^j is measured in the unit of account of state $s, s = 1, \ldots, S$. We say that security j is *real* if it is a promise to deliver the value of a bundle of goods $A_s^j = \left(A_{s1}^j, \ldots, A_{sL}^j\right)$ in each state s so that $V_s^j = p_s A_s^j$. If the first good is chosen as the numeraire for keeping accounts in each state, then $p_{s1} = 1$ and a security j that is real but which only delivers units of the first good is called

a *numeraire* security: for such a security $V_s^j = A_{s1}^j$, $s = 1, \ldots, S$. Security j is said to be *nominal* if its payoff V_s^j is independent of the spot prices p_s. Whatever the type of the security, its price at date 0 is denoted by q_j, and the vector of all security prices is $q = (q_1, \ldots, q_J)$.

Each agent trades on the financial markets choosing a portfolio $z^i = \left(z_1^i, \ldots, z_J^i\right)$ of the securities. These transactions on the financial markets redistribute the agent's income across time and the states. The income acquired or sacrificed at date 0 is $-qz^i = -\sum_{j=1}^{J} q_j z_j^i$(if $z_j^i < 0$, agent i sells security j, i.e. uses security j to borrow; if $z_j^i > 0$, agent i buys security j, i.e. uses security j to save). The income earned or due in state s is $V_s z^i = \sum_{j=1}^{J} V_s^j z_j^i$, where V_s denotes row s of the SxJ matrix V of security payoffs. These income transfers serve to finance the excess expenditures $p_s\left(x_s^i - \omega_s^i\right)$ of the planned consumption stream $x^i = \left(x_0^i, x_1^i \ldots, x_S^i\right)$. Thus the agent's budget set, when current and anticipated prices are (p, q), is given by

$$\mathscr{B}(p, q, \omega^i) = \left\{ x^i \in \mathbb{R}_+^{L(S+1)} \middle| \begin{array}{ll} p_0\left(x_0^i - \omega_0^i\right) = -qz^i, & z^i \in \mathbb{R}^j \\ p_s\left(x_s^i - \omega_s^i\right) = V_s z^i, & s = 1, \ldots, S \end{array} \right\}$$

Each agent i has a preference ordering over the consumption streams $x^i \in \mathbb{R}_+^{L(S+1)}$ which is represented by a utility function $u^i : \mathbb{R}_+^{L(S+1)} \rightarrow \mathbb{R}$ which is typically assumed to have 'nice' properties of strict quasi-concavity, monotonicity and smoothness.

An equilibrium of 'plans, prices, and price expectations' in Radner's (1972) terminology, also called a *spot-financial market equilibrium,* is defined as a pair of actions and prices $((\overline{x},\overline{z}), (\overline{p},\overline{q}))$ such that $(\overline{x}^i, \overline{z}^i)$ maximizes $u^i(x^i)$ over the budget set $\mathscr{B}(\overline{p},\overline{q}, \omega^i)$, $i = 1, \ldots, I$.

the spot markets clear:

$$\sum_{i=1}^{I} \left(\overline{x}_s^i - \omega_s^i\right) = 0, s = 0, \ldots, S$$

the financial markets clear: $\sum_{i=1}^{I} \overline{z}^i = 0$.

The market-clearing conditions (ii) for the agents' planned consumption vectors at date 1 (for $s = 1,\ldots, S$) are what Radner called an equilibrium of 'plans', since the planned consumptions of all agents are compatible, and the anticipated vector of prices p_s for each states s will be an equilibrium vector of spot prices if state s occurs (equilibrium of 'expectations'). Because agents trade at each date this is also called a *sequential equilibrium*.

If the rank of the payoff matrix V is S, so that all possible income transfers from date 0 to date 1 are feasible (at a cost), then we say that financial markets are *complete*. Otherwise, if rank $(V) < S$, then financial markets are *incomplete*, and the corresponding equilibrium is often called a GEI equilibrium.

Existence

If all securities are either numeraire or nominal securities, the budget set $\mathscr{B}(p, q, \omega^i)$ depends continuously on the prices (p, q). To prove existence of equilibrium with such securities, the main insight, over and above the techniques used in classical general equilibrium theory, is that the set of candidate equilibrium prices q for the securities must be restricted to no-arbitrage prices. A portfolio $z \in \mathbb{R}^J$ is called an arbitrage portfolio if $qz \leq 0$ (by selling some securities and buying others the portfolio has no cost) and $Vz \geq 0$ (the date 1 payoff is non-negative) with at least one inequality. An arbitrage portfolio enables an agent to get something (a positive income in some state) without incurring any cost. q is a *no-arbitrage* vector of security prices if it does not admit an arbitrage portfolio.

Much of the modern theory of finance consists in exploring the consequences of no-arbitrage for

the pricing of securities. The analysis centres around the following characterization of no-arbitrage, which is also fundamental for proving existence of an equilibrium in the GEI model. If V is a fixed matrix of payoffs for the securities, q is a no-arbitrage vector of security prices if and only if there exists a strictly positive vector of present-value prices $\pi = (\pi_1, \ldots, \pi_S)$ for income across the states at date 1 such that the price of each security is the present value of its payoff stream: $q_j = \sum_{s=1}^{S} \pi_S V_s^J, s = 1, \ldots, S$. By working with the present-value prices π, the standard Kakutani fixed point theorem can be used to prove existence of an equilibrium with numeraire or nominal securities.

When the securities are real a new difficulty appears, since the payoffs $V_s^j = p_s A_s^j$ depend on the spot prices, so that when the vector of spot prices p changes, the rank of the payoff matrix V can change, leading to discontinuities in the agents' demand functions. As Hart (1975) showed, this can lead to nonexistence of equilibrium. The same kind of discontinuity can appear in the multi-period model with long-lived securities even if the securities are nominal or numeraire. Overcoming this difficulty, i.e. showing that the economies which do not have equilibria are exceptional, has required sophisticated techniques of differential topology which we do not attempt to describe here. The first result was obtained by Duffie and Shafer (1985), and a survey of the methods used for proving generic existence is provided by Magill and Shafer (1991).

Optimality

Since there is a spot market – and thus a price – for each good in each state, the agents' rates of substitution among goods in each state are equalized. To obtain Pareto optimality the additional condition required is that the agents' present-value vectors

$$\pi^i = \left(\pi_s^i\right)_{s=1}^S = \left(\frac{\partial u^i / \partial x_{s1}^i}{\partial u^i / \partial x_{01}^i}\right)_{s=1}^S$$

for income across the states at date 1 (with good 1 as the numeraire) are equalized. The income transfers $(\tau^i)_{i \in I}$ needed for such equalization depend on the risk profiles of the agents' endowments $(\omega^i)_{i \in I}$. Pareto optimality can thus only be expected 'for sure' if any income transfer τ^i is achievable by the choice of a portfolio, i.e. if for any $\tau^i \in \mathbb{R}^S$ there exists $z^i \in \mathbb{R}^J$ such that $Vz^i = \tau^i$. This requires that rank $(V) = S$, namely complete markets. If markets are incomplete, although for particular endowment profiles the necessary income transfers can be achieved through the markets – for instance, the endowments could be Pareto optimal – it can be shown that for almost all endowment profiles $(\omega^i)_{i \in I}$, Pareto optimality is not achievable by a GEI equilibrium.

Since the GEI model involves an imperfection, Pareto optimality is too demanding a criterion. Constrained Pareto optimality, which respects the constraints on the possible income transfers, is a more useful benchmark for judging whether competitive markets lead to the best possible resource allocations given the constraints. An equilibrium allocation is *constrained Pareto optimal* if a 'planner' who can change agents' consumption and portfolios at date 0, but must otherwise let the existing markets induce the allocation at date 1, cannot improve on the allocation. Surprisingly a GEI equilibrium is typically (generically in endowments and preferences) not constrained Pareto optimal. This property was first brought to light by Stiglitz (1982) and formally established by Geanakoplos and Polemarchakis (1986). The channel for the improvement is the change in relative prices at date 1 – the prices $(p_s)_{s=1}^S$ which clear the markets – induced by the change $(dz^i)_{i \in I}$ made by the planner in the agents' date 0 portfolios.

This phenomenon can be seen most simply in a model in which relative prices at date 1 fall out directly as the marginal products of a neo-classical production function. Suppose that at date 0 there is a stock of a single good which can either be consumed or carried over to date 1 as capital input. At date 1 a firm uses this capital and labour to produce a consumption good, with a constant-

returns production function $y = F(K, L)$. Each agent has the same initial endowment ω_0 of the good at date 0 and has a risky labour endowment. At the beginning of date 1, nature draws n agents who are given ℓ_b units of (effective) labour, the remaining $I - n$ agents being given ℓ_g units of labour, with $\ell_b < \ell_g$. There are thus $I!/(n!(I - n)!)$ aggregate states of nature, all equiprobable, which differ from one another by the names of the agents who have the good and the bad draw for their labour endowment. In every state the total supply of labour is the same, $L = n\ell_b + (I - n)\ell_g$. From the point of view of the agents all states in which they have a good draw are equivalent, so that each agent perceives a probability $\rho = n/I$ of having a bad draw and $1 - \rho$ of having a good draw. There are no insurance markets against these labour risks: there is only one security (capital) to transfer income to the two possible outcomes next period, thus markets are incomplete.

Assuming that all agents have the same utility function $U(x_0, x_1) = u(x_0) + \beta E(u(x_1))$ where $0 < \beta \leq 1$ and $x_1 = (x_b, x_g)$, the equilibrium (k, w, R) is characterized by the budget equations, FOCs and market clearing equations

$$x_0 = \omega_0 - k, x_b = w\ell_b + Rk, x_g = w\ell_g + Rk,$$
$$u'(x_0) = \beta\big(\rho u'(x_b) + (1-\rho)u'(x_g)\big)R$$

$$w = F_L(K,L), \quad R = F_K(K,L),$$
$$K = kI, \; L = \big(\rho\ell_b + (1-\rho)\ell_g\big)I$$

Suppose a planner changes the investment k chosen by the typical agent at date 0 by dk; then

$$dx_0 = -dk, dx_b = dw\ell_b + dRk + Rdk,$$
$$dx_g = dw\ell_g + dRk + Rdk$$

which induces a change in the wage and rental rates

$$dw = F_{LK}(K,L)Idk, \quad dR = F_{KK}(K,L)Idk$$

Substituting (dx_0, dx_b, dx_g) the *direct effect* of the change dk is zero in view of the first-order condition for the optimal choice of k, but the *price effects* remain so that

$$dU = \beta\big[\big(\rho u'(x_b)\ell_b + (1-\rho)u'(x_g)\ell_g\big)dw + \big(\rho u'(x_b) + (1-\rho)u'(x_g)\big)kdR\big]$$

Let $\overline{\ell} = \rho\ell_b + (1-\rho)\ell_g$ denote the mean labour endowment. Since F_K is homogenous of degree 0, $dw\overline{\ell} + dRk = 0$. The terms in dw and dR would cancel if $u'(x_b) = u'(x_g)$ i.e. in the case of complete insurance markets. In the absence of insurance markets, $u'(x_b) \neq u'(x_g)$ and $dU \neq 0$. dU can be written as

$$\begin{aligned} dU &= \beta(E(u'(x_1)\ell_1)dw + E(u'(x_1)kdR) \\ &= \beta(E(u'(x_1)E(\ell_1)dw + cov(u'(x_1,\ell_1)dw \\ &\quad + E(u'(x_1))kdR) \\ &= \beta cov(u'(x_1),\ell_1)dw \end{aligned}$$

Since u' is decreasing, it follows that $cov(u'(x_1), \ell_1) < 0$. A change $dk < 0$, which implies $dw < 0$, leads to an increase in welfare, $dU > 0$.

Reducing saving at date 0 increases date 0 consumption and reduces consumption at date 1, and to terms of first order, the direct effect of the change in consumption is zero, since agents have optimized on their choice of saving at equilibrium. But the price of capital increases and the price of labour decreases, shifting the representative agent's income away from the risky labour income ($w\ell_b$, $w\ell_g$) and towards the sure return (kR, kR) on capital. The price effect reduces the variability of date 1 consumption, improving the welfare of the representative agent.

The change in prices (partially) replaces the insurance market which is missing.

A reduction dk in the agents' savings can also be achieved if the planner imposes a tax t on saving and redistributes the proceeds lump sum ($T = kt$) to the agents. The property of constrained suboptimality of a GEI equilibrium suggests that appropriate taxes on securities could be used to improve on the allocation achieved with incomplete markets. However Citanna et al. (2006) have shown that to be sure to achieve a Pareto improvement in this way the number of securities (J) must exceed the number of agents (I), since the needed reallocations can

only be achieved for sure if there are as many instruments (taxes) as objectives (agents' utilities).

Determinacy

The study of determinacy of equilibrium has served to uncover important differences between economies in which securities are nominal and those in which they are real. The study of economies with nominal securities led to the realization that monetary considerations need to be incorporated as an integral part of the sequential model.

In an economy in which securities are real, since the payoff of each security is proportional to the spot prices, doubling spot prices in a state doubles the payoffs of the securities, leaving agents' budget sets unchanged $\left(p_s\left(x_s^i - \omega_s^i\right) = \sum_{j=1}^{J} p_s A_s^j z^i\right)$. Thus price levels do not matter, and as in the standard GE model, the spot prices in each state can be normalized (e.g. $p_{s1} = 1, s = 0, \ldots, S$). Using arguments of differential topology analogous to those developed for the GE model it can be shown that generically (in endowments) an economy has only a finite number of equilibria – in short, with real assets GEI allocations are determinate.

When the securities are nominal, since the payoffs are independent of the spot prices, price levels matter. Doubling the price level in state s halves the purchasing power of the income promised by the assets in this state $\left(p_s\left(x_s^i - \omega_s^i\right) = \sum_{j=1}^{J} V_s^j z^i\right)$. This reasoning is insufficient to conclude that agents will be affected: for if agents correctly anticipate the 'doubling of the price level in state s' then they may adapt their portfolios accordingly to annul the effect. This is where the incompleteness of the security structure enters the picture. If the financial markets are complete, any change in the price levels across the states at date 1 can be 'undone' by a corresponding change in the portfolio chosen at date 0, so that again equilibrium allocations do not depend on price levels. If markets are incomplete, some changes in price levels cannot be 'undone' by changes in the agents' choices of portfolios, so that equilibrium allocations are different with different price levels. Thus if the security structure consists of nominal securities and is incomplete, and if a GEI equilibrium is defined by conditions (i), (ii) and (iii) above so that nothing determines price levels, then there is a continuum of equilibrium outcomes. This property was first noted by Cass (1989), and the precise characterization of the dimension of the manifold of equilibria was studied by Balasko and Cass (1989) and Geanakoplos and Mas-Colell (1989).

Magill and Quinzii (1992) argued that a nominal contract is a promise to make a deferred payment of a sum of money, and that such promises only come to be made in an economy in which money is already used as a medium of exchange and a unit of account. What is needed therefore is a way of introducing money as a medium of exchange in the GEI model so that price levels are determined by the monetary side of the economy. They introduce a highly stylized (some might say 'brute force') model in which Clower's (1967) idea that only money can buy goods leads to a system of $S + 1$ quantity theory equations

$$\sum_{i=1}^{I} p_s x_s^i = M_s, \quad s = 0, 1, \ldots, S$$

asserting that the demand for money for transactions must equal the supply of money M_s in each state – the vector $M = (M_0, M_1, \ldots, M_S)$ defining the monetary policy. When agents correctly anticipate the monetary policy and markets are complete, monetary policy does not affect the equilibrium allocation – a change from M to M' just leads to a change of portfolios financing the same allocation – but if markets are incomplete, different monetary policies lead to different allocations. The indeterminacy in the GEI model without price level determination becomes the property that, with nominal assets and incomplete markets, correctly anticipated monetary policy has real effects.

The need to introduce money explicitly into the GEI model with nominal assets has prompted the development of monetary models which are

closer to the cash-in-advance models of macroeconomics, in which the interest cost of holding money (seignorage tax) is explicitly modelled. This has led to interesting ways of examining the structure of a monetary equilibrium model over a finite horizon (Dubey and Geanakoplos 2003; Drèze and Polemarchakis 2000) and to exploring the conditions (nonRicardian versus Ricardian) under which monetary and fiscal policies do or do not determine price levels (Nakajima and Polemarchakis 2005).

In this short entry we have focused on properties of the GEI model in the simplest two-period or finite-horizon exchange setting. A more complete analysis of this model can be found in Magill and Quinzii (1996). A host of interesting new issues arise when the model is extended to an infinite horizon and in addition is extended to incorporate default, which bring to light the close connexion between GEI and macroeconomics.

See Also

- ▶ Arrow–Debreu Model of General Equilibrium
- ▶ Computation of General Equilibria
- ▶ Computation of General Equilibria (New Developments)
- ▶ General Equilibrium

Bibliography

Arrow, K. 1953. Le role des valeurs boursières pour la répartition la meilleure des risques. *Econométrie, Collogues Internationaux du Centre National de la Recherche Scientifique* 40: 41–47; English version: The role of securities in the optimal allocation of risk bearing. *Review of Economic Studies* (1964) 31: 91–96.

Balasko, Y., and D. Cass. 1989. The structure of financial equilibrium with exogenous yields: The case of incomplete markets. *Econometrica* 57: 135–162.

Cass, D. 1989. Sunspots and incomplete financial markets: The leading example. In *The economics of imperfect competition Joan Robinson and beyond*, ed. G. Feiwe. London: Macmillan.

Citanna, A., H.M. Polemarchakis, and M. Tirelli. 2006. The taxation of trades in assets. *Journal of Economic Theory* 126: 299–313.

Clower, R.W. 1967. A reconsideration of the microfoundations of monetary theory. *Western Economic Journal* 6: 1–8.

Debreu, G. 1959. *Theory of value*. New York: Wiley.

Drèze, J.H., and H. Polemarchakis. 2000. Monetary equilibria. In *Economic essays: A Festschift for Werner Hildenbrand*, ed. G. Debreu, W. Neufeind, and W. Trocke, 83–108. Heidelberg/New York: Springer.

Dubey, P., and J. Geanakoplos. 2003. Inside and outside fiat money, gains to trade and IS-LM. *EconomicTheory* 21: 347–397.

Duffie, D., and W. Shafer. 1985. Equilibrium in incomplete markets, I: A basic model of generic existence. *Journal of Mathematical Economics* 14: 285–300.

Geanakoplos, J., and A. Mas-Colell. 1989. Real indeterminacy with financial assets. *Journal of Economic Theory* 47: 22–38.

Geanakoplos, J., and H. Polemarchakis. 1986. Existence, regularity, and constrained suboptimality of competitive allocations when markets are incomplete. In *Uncertainty, information and communication: Essays in honor of Kenneth Arrow*, vol. 3, ed. W.P. Heller, R.M. Ross, and D.A. Starrett. Cambridge: Cambridge University Press.

Hart, O.D. 1975. On the optimality of equilibrium when the market structure is incomplete. *Journal of Economic Theory* 11: 418–443.

Magill, M., and M. Quinzii. 1992. Real effects of money in general equilibrium. *Journal of Mathematical Economics* 21: 301–342.

Magill, M., and M. Quinzii. 1996. *Theory of incomplete markets*. Boston: MIT Press.

Magill, M., and W. Shafer. 1991. Incomplete markets. In *Handbook of mathematical economics*, vol. IV, ed. W. Hildenbrand and H. Sonnenschein. Amsterdam: North Holland.

Nakajima, T., and H. Polemarchakis. 2005. Money and prices under uncertainty. *Review of Economic Studies* 72: 223–246.

Radner, R. 1972. Existence of equilibrium of plans, prices and price expectations in a sequence of markets. *Econometrica* 40: 289–304.

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

Stiglitz, J.E. 1982. The inefficiency of stock market equilibrium. *Review of Economic Studies* 49: 241–261.

General Purpose Technologies

Peter L. Rousseau

Abstract

Economists have come to use the term ‘general purpose technology’ (GPT) to describe technological advances that pervade many sectors, improve rapidly, and spawn further

innovations. This article addresses the concept of a GPT by example, showing the extent to which electricity and information technology might qualify as members of this special class of inventions, as opposed to more ordinary ones.

Keywords

Diffusion of technology; Electricity; General purpose technologies; Growth, models of; Information technology and economic growth; Innovation; Intertemporal substitution; Patents; Productivity growth; Skill premium; Technical change

JEL Classifications

O4

Economists have long been interested in how technological change affects long-run growth and aggregate fluctuations, yet it remains most often treated as incremental in nature, adding only a trend to standard growth models. History tells us, however, that such change can appear in bursts, with flurries of innovative activity following the introduction of a new core technology. This observation leads economists to reserve the term 'general-purpose technology' (GPT) to describe fundamental advances that drive these flurries, which in turn transform both household life and the ways in which firms conduct business. Over the past 200 years or so, steam, electricity, internal combustion, and information technology (IT) seem to have served as GPT-type technologies. They affected entire economies. Earlier, the very ability to communicate in writing and later to disseminate written information via the printed page also appears to fit well into the idea of a GPT.

The notions that GPTs differ from the more incremental refinements that occur in between their arrivals and that they represent real-side shocks that permanently change the nature of production and preferences provide the basis of a potentially useful way to organize thinking about long-run economic fluctuations and growth. But to support such a view with anything more than casual observation, it is necessary to establish criteria for determining just what features a technology must possess in order to be a GPT rather than a more ordinary invention. This article defines GPTs in terms of a number of tangible criteria, and then uses two candidate GPTs, electrification and IT, to demonstrate how identification of a GPT might proceed. Attention then turns to other indicators that may signal the start of a GPT era.

Dating a GPT's Arrival

Associating a point in time with a GPT's 'arrival' depends on what exactly one means by this term. If defined with a measure such as, in the case of electrification, attaining a one per cent share of horsepower in the manufacturing sector, then some time around 1895 might be appropriate. This coincides roughly with the start-up of the world's first large scale hydroelectric power facility at Niagara Falls, New York, in 1894. It would be reasonable to argue, however, that electricity arrived earlier, perhaps in 1882 when Thomas Edison brought the first centralized electricity system online at the Pearl Street station in lower Manhattan. For IT, it is true that mainframe computers had existed for two decades before the invention of the 4004 chip in 1971, and had even been used to project the winner of the 1952 US presidential election. Yet, if measured by the attainment of a one per cent share in the industrial sector's stock of equipment, 1971 remains the most likely candidate for dating IT's 'arrival'.

Whether electricity and IT arrived in 1895 and 1971, respectively, or some time prior to these dates, one characteristic noted by David (1991) is that neither delivered productivity gains immediately. Indeed, productivity growth as measured by output per man-hour seems to have been relatively high in the 1870s, when steam was the dominant power source for industry, but fell as electrification arrived in the 1880s and 1890s. It was only in the period after 1915, which also saw the diffusion of secondary motors and the widespread establishment of centralized power distribution systems, that measured productivity numbers began to rise. (This can be seen in the

series for output per man-hour in the non-farm business sector from US Census Bureau, 1975, Series D684, p. 162.) Further, Intel's 1971 invention of the 4004 microprocessor (the key component in the first generation of personal computers), if taken to be the start of the IT era, did not reverse the decline in productivity growth that had begun more than a decade earlier.

Identification of a New Core Technology as a GPT

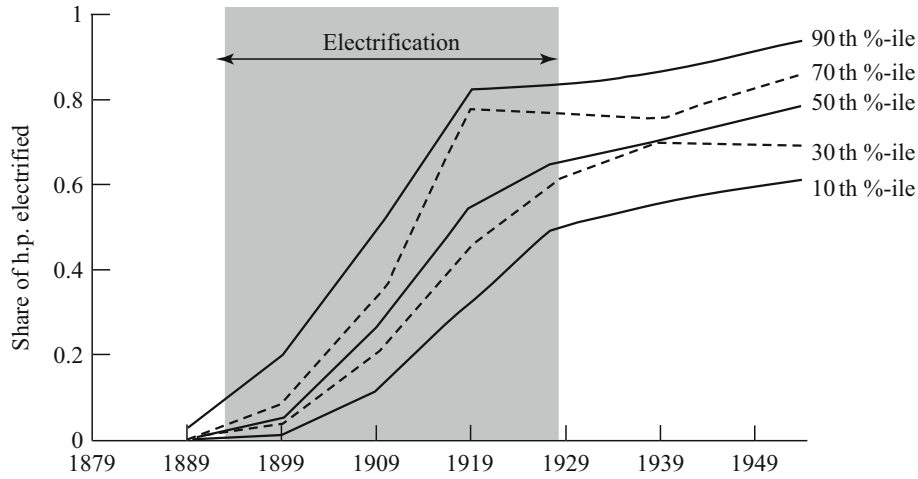
Once the arrival date of a new technology has been established, identification of that technology as a GPT can proceed by considering characteristics associated with its diffusion. One set of criteria, proposed by Bresnahan and Trajtenberg (1995), suggests that a GPT should have the following three characteristics:

1. *Pervasiveness:* the GPT should spread to most sectors.
2. *Improvement:* the GPT should get better over time and, hence, should keep lowering the costs of its users.
3. *Innovation spawning:* the GPT should make it easier to invent and produce new products or processes.

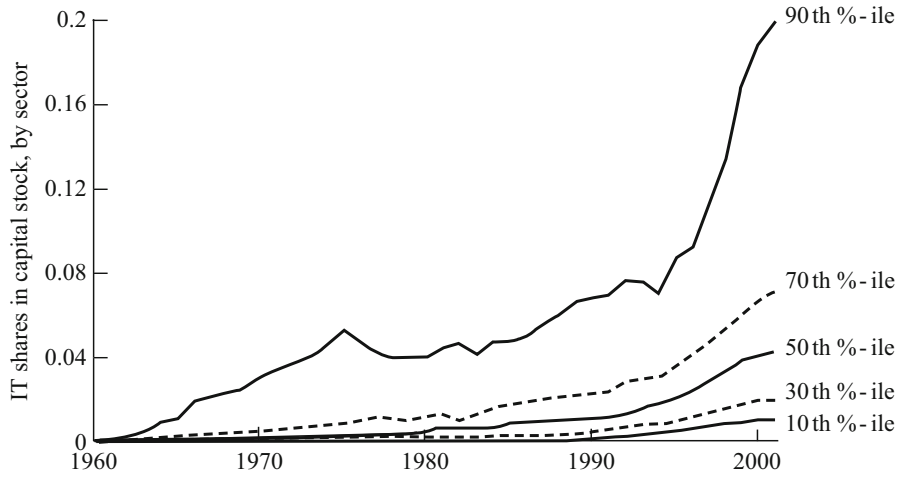
Most technologies possess each of these characteristics to some degree, and therefore a GPT cannot differ qualitatively from them. But the extent to which technologies have all three characteristics should determine which ones are likely to be GPTs.

For example, both electrification and IT were pervasive, and so might qualify as GPTs under the first criterion, yet had quite different absorption paths across sectors. Figure 1 shows the shares of total horsepower electrified in manufacturing sectors at ten-year intervals from 1889 to 1954 in percentile form, with the shaded area highlighting the period of electricity's most rapid diffusion. Figure 2 shows the spread of IT, measured as the share of IT equipment in the capital stock at the two-digit standard industry classification level. The striking difference between the two figures is that electricity diffused uniformly across sectors while the adoption of IT was not as widespread. On this count, then, electricity would be the stronger GPT candidate.

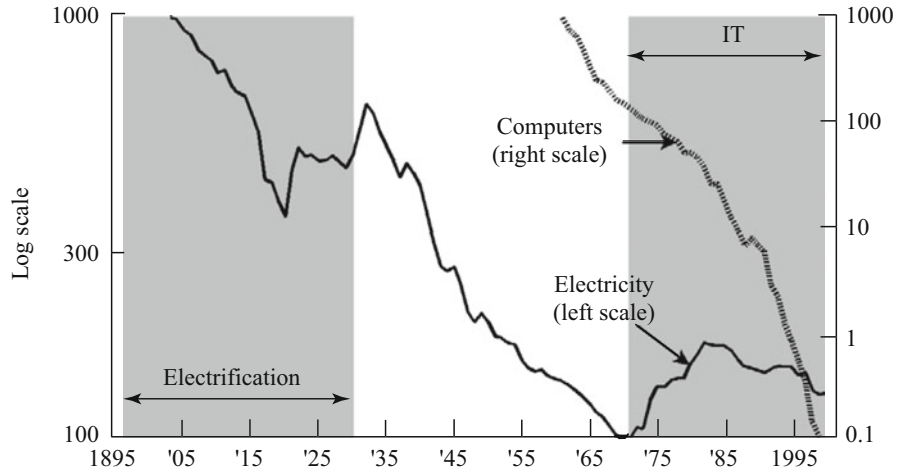
Presumably, the second characteristic – improvement – would show up in a decline in prices associated with the technology, an increase in quality, or both. How much a GPT improves can therefore be measured by how much cheaper a unit of quality gets over time. If the new technology is embodied in capital and begins to account



General Purpose Technologies, Fig. 1 Shares of electrified horsepower by manufacturing sector in percentiles, 1890–1954 (*Source*: DuBoff (1964, Tables E-ll and E-12–12e))



General Purpose Technologies, Fig. 2 Shares of IT equipment and software in the capital stock by sector in percentiles, 1960–2001 (*Source*: Detailed non-residential fixed asset tables in fixed 1996 dollars made available by the US Bureau of Economic Analysis (2004))



General Purpose Technologies, Fig. 3 Price indices for products of two 'GPT eras', 1895–2000. *Sources*: The quality-adjusted price index for IT is formed by joining the 'final' price index for computer systems from Gordon (1990, Table 6.10, col. 5, p. 226) for 1960–78 with the pooled index developed for desktop and mobile personal computers by Berndt et al. (2000, Table 2, col. 1, p. 22) for 1979–99. Electricity prices are averages of all electric energy services in cents per kilowatt hour from US Census Bureau (1975, series S119, p. 827) for 1903, 1907, 1917, 1922, and 1926–70, and from the US Census Bureau, *Statistical Abstract of the United States*, for 1971–89. For 1990–2000, prices are US city averages (June figures) from the US Bureau of Labor Statistics. Both indices are set to 1,000 in the first years of the samples (that is, 1903 and 1960)

for an increasing share of the net capital stock, capital should on the whole be getting cheaper faster during a GPT era, but especially capital that is tied to the new technology.

Figure 3 plots the price of the components of the aggregate capital stock tied to the two GPTs. Because deflators for electrically powered capital are not available in the first half of the 20th century, the figure compares the declines in relative price of electricity itself with the quality-adjusted price of computers, both relative to the consumption price index. The use of the left-hand scale for

electricity and the right-hand scale for computers underscores the extraordinary decline in computer prices since 1960 relative to electricity. While electricity prices fall by a factor of 10, the computer price index falls by a factor of 10,000!

It can be said that the electricity index, being the price of a kilowatt hour, understates the accompanying technological change because it does not account for improvements in electrical equipment, and especially improvements in the efficiency of electrical motors. Based on the price evidence in Fig. 3, however, both electricity and computers might qualify as GPTs, with computers clearly more revolutionary.

With respect to the ability to generate further innovation, it is reasonable to assume that any GPT will affect all sorts of production processes, including those for invention and innovation. Some GPTs will be biased towards helping to produce existing products, others towards inventing and implementing new ones. Electricity and IT have both helped reduce the costs of making existing products, and they both spawn innovation. The 1920s especially saw a wave a new products powered by electricity, and the computer is now embodied in many new products as well. But the evidence suggests that IT has contributed more to furthering innovation.

In particular, patenting should be more intense after a GPT arrives and while it is spreading due to the introduction of related new products. US patent data confirm this, showing two surges in the annual number of invention patents issued per capita from 1890 to 2000 – one between 1900 and 1930, and the other after 1977. At the same time, the surge during the IT period was stronger than that observed during electrification. Interestingly, the slow rate of patenting during the Second World War years and the acceleration immediately thereafter suggests that there is some degree of intertemporal substitution in the release of new ideas away from times when they might be more difficult to popularize and towards times better suited for the entry of new products.

Of course, patent data may reflect fluctuations in the number of actual inventions or may simply reflect changes in the law that raise the propensity to patent. The distinction is important because, over longer periods of time, patents may reflect policy rather than invention. Kortum and Lerner (1998) analyse this question and find that the surge of the 1990s was worldwide, but not systematically related to country-specific policy changes, and they conclude that technology was the cause of the surge.

Other Characteristics of GPTs

In addition to the three basic qualities of a GPT, there are other, less direct signals implied by various theoretical models that deal with GPTs. These models predict the following:

1. *New ideas should come to market faster*. If a new technology has the potential for large productivity gains, firms will spend less time perfecting ideas associated with the new technology in order to realize the gains sooner (see, for example, Jovanovic and Rousseau, 2001).
2. *Entry, exit and mergers should rise.* New technologies may require some relocation of assets from firms that are unable to adopt them effectively to others with managements better equipped for their deployment (see, for example, Jovanovic and Rousseau, 2002).
3. *Young and small firms should do better.* The ideas and products associated with the GPT will often be brought to market by new firms. The market share and market value of young firms should therefore rise relative to old firms.
4. *Stock prices should initially fall*. The value of old capital should fall in anticipation of the new and more productive technology. How fast it falls depends on the way that the market learns of the GPT's arrival (see, for example, Hobijn and Jovanovic, 2001).
5. *Interest rates and the trade deficit should be affected*. The rise in desired consumption relative to output should cause interest rates to rise or the trade balance to worsen.
6. *The skill premium should rise.* If the GPT is not user-friendly at first, skilled people will be in greater demand when the new technology

arrives, and their earnings should rise compared with those of the unskilled.

The available evidence suggests that predictions (1)–(3) hold for both the electrification and IT eras, but that a stock market decline (4) occurred only at the start of the IT period. Interest rates (5) rose in both eras, but the electrification period was associated with a trade surplus due to the First World War. It also appears that the skill premium (6) has risen over the IT period, but evidence of a rise in the electrification era is weaker.

To sum up, based upon the criteria chosen and the available evidence, both electricity and IT were pervasive, improving, and innovation-spawning, and thus seem to qualify as GPTs. At the same time, electricity was more pervasive, affecting sectors faster and more evenly than IT, while IT improved more dramatically, with computer prices falling more than 100 times faster than the price of electricity. IT also seems to have generated more innovation than electricity, and the initial productivity slowdown was also deeper in the IT era. All this would lead one to regard IT as the more 'revolutionary' GPT.

This is not to say that the differences between electrification and IT, or indeed between any two candidate GPTs, are unimportant. At the same time, the GPT paradigm emphasizes the commonalities, namely, that technological progress is uneven, that it does entail the episodic arrival of new core technologies, and that these GPTs bring on turbulence and lower growth early on and higher growth and prosperity later. Interestingly, the IT era has already outlasted that of electrification, but even six decades after what Field (2003) has called the 'most technologically progressive decade of the century' (that is, the 1930s), electricity has yet to become obsolete. Given the multitude of firms and households that have not quite yet adopted IT, its continuing price decline and the widespread increases in computer literacy among children and adults worldwide suggest that perhaps the most productive period of this GPT still lies ahead.

See Also

- ▶ Diffusion of Technology
- ▶ Electricity Markets
- ▶ Information Technology and the World Economy
- ▶ Technical Change

Acknowledgments The author acknowledges financial support from the National Science Foundation.

Bibliography

Berndt, E.R., E.R. Dulberger, and N.J. Rappaport. 2000. *Price and quality of desktop and mobile personal computers: A quarter century of history.* Working paper. Cambridge, MA: Sloan School of Management, MIT.

Bresnahan, T.F., and M. Trajtenberg. 1995. General purpose technologies: Engines of growth? *Journal of Econometrics* 65: 83–108.

David, P.A. 1991. Computer and dynamo: The modern productivity paradox in a not- too-distant mirror. In *Technology and productivity: The challenge for economic policy.* Paris: OECD.

DuBoff, R.B. 1964. Electric power in American manufacturing, 1889–1958. Ph.D. thesis, University of Pennsylvania.

Field, A. 2003. The most technologically progressive decade of the century. *American Economic Review* 93: 1399–1414.

Gordon, R.J. 1990. *The measurement of durable goods prices*. Chicago: University of Chicago Press.

Hobijn, B., and B. Jovanovic. 2001. The IT revolution and the stock market: Evidence. *American Economic Review* 91: 1203–1220.

Jovanovic, B., and P.L. Rousseau. 2001. Why wait? A century of life before IPO. *American Economic Review Papers and Proceedings* 91: 336–341.

Jovanovic, B., and P.L. Rousseau. 2002. The Q-theory of mergers. *American Economic Review Papers and Proceedings* 92: 198–204.

Kortum, S., and J. Lerner. 1998. Stronger protection or technological revolution: What is behind the recent surge in patenting? *Carnegie-Rochester Conference Series on Public Policy* 48: 247–304.

US Bureau of Economic Analysis. 2004. *Survey of current business*. Washington, DC: Government Printing Office.

US Census Bureau. 1975. *Historical statistics of the United States, Colonial Times to 1970*. Washington, DC: Government Printing Office.

US Census Bureau (various years). *Statistical abstract of the United States*. Washington, DC: Government Printing Office.

G

General Systems Theory

Kenneth E. Boulding

The term 'general systems' refers to a movement among a wide variety of scholars to overcome the barriers of communication which divide the established disciplines, by developing theoretical concepts and systems which are common to the different disciplines. Biologist Ludwig von Bertalanffy originated the movement with his concept of 'open systems'. The Society for General Systems Research, originally called the Society for the Advancement of General Systems, was founded at a meeting at the American Association for the Advancement of Science in Berkeley, California, in December 1954. The economist Kenneth E. Boulding was the first president. The Society issues the *General Systems Yearbook*, partly of reprinted, partly of original articles, of which the first editor was Anatol Rapoport, a mathematician and game theorist. The yearbooks are still published, and a number of journals now contribute to the field.

In Europe general systems has frequently been identified with 'cybernetics', originated by Norbert Wiener of MIT in 1948, which is the study of both the equilibrium and disequilibrium systems which involve feedback. A thermostat is a good example of an equilibrium system with negative feedback. The equilibrium is manipulable. It is the temperature at which the thermostat is set. If the temperature rises above this, the thermostat turns the furnace off; if it falls below it, the thermostat turns the furnace on. All such cybernetic systems, of which there are many, such as homoeostatic mechanisms of the body, exhibit cycles, the period and magnitude of which depend mainly on the time of response of the feedback. The tendency of the market price system of relative prices to fluctuate around an equilibrium and the tendency of competitive markets in commodities and securities to fluctuate in aggregate or average prices is a good example of negative feedback provided by the behavioural reactions to price above or below what is regarded as normal. Inflation is frequently an example of positive feedback, especially hyperinflation, where a rise in the price level produces both expectations which lead to a continued rise and also a partial collapse of the tax system and budget deficits, which likewise feed the continuing rise. Deflation, such as occurred during the Great Depression of 1929–33, is also a positive-feedback process, in which, for instance, declining profits produce declining investment, which produces further declining profits, further declining investments, and so on.

Another important line of development of general systems has been the development of a general theory of the ontogeny, structure and behaviour of organisms, ranging from the cell, the organ, the living organism, the group, the social organization, the nation-state, and so on. James Grier Miller has made important contributions to this, particularly in regard to the taxonomy of organisms, and has identified at least 19 necessary components of such structures, common to all of them.

The structure of organisms and organizations is also influenced by the principle of allometry, developed especially by von Bertalanffy, but going back to D'Arcy Thompson in his work *On Growth and Form*. This is the principle that an increase in the linear dimensions of any structure, keeping the same proportions, will increase the areas as a square and the volumes as the cube of the linear increase. Thus a two-inch cube has four times the area and eight times the volume of a one-inch cube. This explains why structure is a function of scale. Some properties depend on the linear dimensions, like the transmission of information or fluids. Some depend on the areas, like chemical exchange and structural strength; some depend on the volumes, like weight or mass. This goes a long way to explaining why structure changes with size and must change with growth. This principle also applies to social organizations and firms in terms of hierarchy, specialization, diffusion of responsibility, and so on. In economics it is responsible for such phenomena as diseconomies of scale unless there is structural

change and also for the less well recognized phenomenon of organizational failure when growth takes place without adequate structural change.

Another aspect of general systems is the development of more general ecological and evolutionary theory. There are many parallels between ecological theory in biology and the theory of a general equilibrium and development of commodities. One may claim Adam Smith, indeed, as perhaps the first ecological and evolutionary theorist, perceiving the economic system to be an ecosystem of commodities with equilibrium populations at which births (production) and deaths (consumption) are equal, with the equilibrium population of each commodity being a function of the population of all others. This gives us a system of n equations and n unknowns, as developed, for instance, in economics by Walras. Adam Smith further recognizes what today would be called 'mutation'; that is, changes in the parameters of the system, leading to new equilibrium positions continually as changes take place in the genetic factors in the production of commodities. Adam Smith also recognized that these genetic factors primarily involved changes in human knowledge as a result of a learning process. The classical economist also had something like a food chain theory of the ecosystem of commodities, with the input of food into the food-producer producing a surplus of food which could then feed the producers of other commodities. The main difference between biological and economic systems is that the genetic structure for biological products is contained in the products themselves, whereas in the case of commodities the genetic structure is contained in many other human minds and human artifacts. The economic system is multi-parental. These ideas are not very widely accepted by economists, who still cling to a somewhat Newtonian view of the system, perhaps because they do not conform easily to quantification and mathematization, and emphasize that the real world consists of structure rather than number.

General systems has not established itself well in the role structure of universities; very few have formal programmes in it. The Society for General Systems Research, however, is still very much alive. Other aspects of general systems, such as theories of autopoiesis – that is, the instability of chaos and the spontaneous formation of structures – may turn out to have considerable relevance to economic problems like entrepreneurship and innovation. It cannot be claimed that general systems has had much impact on economics to date, but it is not much more than 30 years old.

See Also

- ▶ Bertalanffy, Ludwig von (1901–1972)
- ▶ Boulding, Kenneth Ewart (Born 1910)
- ▶ Stability

Generalized Method of Moments Estimation

Lars Peter Hansen

Abstract

Generalized method of moments estimates econometric models without requiring a full statistical specification. One starts with a set of moment restrictions that depend on data and an unknown parameter vector to be estimated. When there are more moment restrictions than underlying parameters, there is family of such estimators. The tractable form of the large sample properties of this family facilitates efficient estimation and statistical testing. This article motivates the method, presents some of the underlying statistical properties, and discusses implementation.

Keywords

Calibration; Central limit theorems; Gauss–Markov theorem; Generalized method of moments; Identification; Instrumental variables; Lagrange multipliers; Law of large numbers; Likelihood; Martingales; Maximum

likelihood; Rational expectations models; Sequential estimation; Statistical inference; Stochastic discount factor models; Wald test

JEL Classifications
C10

Introduction

Generalized method of moments (GMM) refers to a class of estimators constructed from the sample moment counterparts of population moment conditions (sometimes known as orthogonality conditions) of the data generating model. GMM estimators have become widely used, for the following reasons:

1. GMM estimators have large sample properties that are easy to characterize. A family of such estimators can be studied simultaneously in ways that make asymptotic efficiency comparisons easy. The method also provides a natural way to construct tests which take account of both sampling and estimation error.
2. In practice, researchers find it useful that GMM estimators may be constructed without specifying the full data generating process (which would be required to write down the maximum likelihood estimator). This characteristic has been exploited in analysing partially specified economic models, studying potentially misspecified dynamic models designed to match target moments, and constructing stochastic discount factor models that link asset pricing to sources of macroeconomic risk.

Books with good discussions of GMM estimation with a wide array of applications include: Cochrane (2001), Arellano (2003), Hall (2005), and Singleton (2006). For a theoretical treatment of this method see Hansen (1982) along with the self-contained discussions in the books. See also Ogaki (1993) for a general discussion of GMM estimation and applications, and see Hansen (2001) for a complementary article that, among other things, links GMM estimation to related literatures in statistics. For a collection of recent methodological advances related to GMM estimation, see the journal issue edited by Ghysels and Hall (2002). While some of these other references explore the range of substantive applications, in what follows we focus more on the methodology.

Set-Up

As we will see, formally there are two alternative ways to specify GMM estimators, but they have a common starting point. Data are a finite number of realizations of the process $\{x_t: t = 1, 2,\ldots\}$. The model is specified as a vector of moment conditions:

$$Ef(x_t, \beta_0) = 0$$

where f has r coordinates and β_0 is an unknown vector in a parameter space $\mathbb{P} \subset \mathbb{R}^k$. To achieve identification we assume that on the parameter space $\mathbb{P}$

$$Ef(x_t, \beta) = 0 \text{ if, and only if } \beta = \beta_0. \quad (1)$$

The parameter β_0 is typically not sufficient to write down a likelihood function. Other parameters are needed to specify fully the probability model that underlies the data generation. In other words, the model is only partially specified.

Examples include:

(a) Linear and nonlinear versions of instrumental variables estimators as in Sargan (1958, 1959), and Amemiya (1974)
(b) Rational expectations models as in Hansen and Singleton (1982), Cumby et al. (1983), and Hayashi and Sims (1983)
(c) Security market pricing of aggregate risks as described, for example, by Cochrane (2001), Singleton (2006), and Hansen et al. (2007)
(d) Matching and testing target moments of possibly misspecified models as described by, for example, Christiano and Eichenbaum (1992), and Hansen and Heckman (1996)

Regarding example (a), many related methods have been developed for estimating correctly specified models, dating back to some of the original applications in statistics of method-of-moments-type estimators. The motivation for such methods was computational. See Hansen (2001) for a discussion of this literature and how it relates to GMM estimation. With advances in numerical methods, the fully efficient maximum likelihood method and Bayesian counterparts have become much more tractable. On the other hand, there continues to be an interest in the study of dynamic stochastic economic models that are misspecified because of their purposeful simplicity. Thus moment matching remains an interesting application for the methods described here. Testing target moments remains valuable even when maximum likelihood estimation is possible (for example, see Bontemps and Meddahi 2005).

Central Limit Theory and Martingale Approximation

The parameter dependent average

$$g_N(\beta) = \frac{1}{N}\sum_{t=1}^{N} f(x_t, \beta)$$

is featured in the construction of estimators and tests. When the law of large numbers is applicable, this average converges to the $Ef(x_t; \beta)$. As a refinement of the identification condition:

$$\sqrt{N} g_N(\beta_0) \Rightarrow \text{Normal}(0, V) \tag{2}$$

where $\Rightarrow$ denotes convergence in distribution and V is a covariance matrix assumed to be nonsingular. In an iid data setting, V is the covariance matrix of the random vector $f(x_t; \beta_0)$. In a time series setting:

$$V = \lim_{N\to\infty} NE\left[g_N(\beta_0) g_N(\beta_0)'\right], \tag{3}$$

which is the long-run counterpart to a covariance matrix.

Central limit theory for time series is typically built on martingale approximation (see Gordin 1969; Hall and Heyde 1980). For many time series models, the martingale approximators can be constructed directly and there is specific structure to the V matrix. A leading example is when $f(x_t; \beta_0)$ defines a conditional moment restriction. Suppose that x_t, $t = 0,1,\ldots$, generates a sigma algebra $\mathcal{F}_t$, $E[|f(x_t; \beta_0)|^2] < \infty$ and

$$E[f(x_{t+\ell}, \beta_0) | \mathcal{F}_t] = 0$$

for some $\text{ø} \geq 1$. This restriction is satisfied in models of multi-period security market pricing and in models that restrict multi-period forecasting. If $\text{ø} = 1$, then g_N is itself a martingale; but when $\text{ø} > 1$, it is straightforward to find a martingale m_N with stationary increments and finite second moments such that

$$\lim_{N\to\infty} E\left[|g_N(\beta_0) - m_N(\beta_0)|^2\right] = 0,$$

where $|\cdot|$ is the standard Euclidean norm. Moreover, the lag structure may be exploited to show that the limit in (3) is

$$V = \sum_{j=-\ell+1}^{\ell-1} E\left[f(x_t, \beta_0) f\left(x_{t+j}, \beta_0\right)'\right]. \tag{4}$$

(The sample counterpart to this formula is not guaranteed to be positive semidefinite. There are a variety of ways to exploit this dependence structure in estimation in constructing a positive semidefinite estimate. See Eichenbaum et al. 1988, for an example.) When there is no exploitable structure to the martingale approximator, the matrix V is the spectral density at frequency zero.

$$V = \sum_{j=-\infty}^{\infty} E\left[f(x_t, \beta_0) f\left(x_{t+j}, \beta_0\right)'\right].$$

Minimizing a Quadratic Form

One approach for constructing a GMM estimator is to minimize the quadratic form:

$$b_N = \arg\min_{\beta \in \mathbb{P}} g_N(\beta)' W g_N(\beta)$$

for some positive definite *weighting matrix* W. Alternative weighting matrices W are associated with alternative estimators. Part of the justification for this approach is that

$$\beta_0 = \arg\min_{\beta \in \mathbb{P}} Ef(x_t, \beta)' W Ef(x_t, \beta).$$

The GMM estimator mimics this identification scheme by using a sample counterpart.

There are a variety of ways to prove consistency of GMM estimators. Hansen (1982) established a uniform law of large numbers for random functions when the data generation is stationary and ergodic. This uniformity is applied to show that

$$\sup_{\beta \in \mathbb{P}} \mid g_N(\beta) - E[f(x_t, \beta)] \mid = 0$$

and presumes a compact parameter space. The uniformity in the approximation carries over directly the GMM criterion function $g_N(\beta)' W g_N(\beta)$. See Newey and McFadden (1994) for a more complete catalogue of approaches of this type.

The compactness of the parameter space is often not ignored in applications, and this commonly invoked result is therefore less useful than it might seem. Instead, the compactness restriction is a substitute for checking behaviour of the approximating function far away from β_0 to make sure that spurious optimizers are not induced by approximation error. This tail behaviour can be important in practice, so a direct investigation of it can be fruitful. For models with parameter separation:

$$f(x, \beta) = Xh(\beta)$$

where X is an $r \times m$ matrix constructed from x and h is a one-to-one function mapping $\mathbb{P}$ into subset of $\mathbb{R}_m$, there is an alternative way to establish consistency (see Hansen 1982 for details). Models that are either linear in the variables or models based on matching moments that are nonlinear functions of the underlying parameters can be written in this separable form.

The choice of $W = V^{-1}$ receives special attention, in part because

$$N g_N(\beta)' V^{-1} g_N(\beta) \Rightarrow \chi^2(r).$$

While the matrix V is typically not known, it can be replaced by a consistent estimator without altering the large sample properties of b_N. When using martingale approximation, the implied structure of V can often be exploited as in formula (4). When there is no such exploitable structure, the method of Newey and West (1987b) and others can be employed that are based on frequency-domain methods for time series data.

For asset pricing models there are other choices of a weighting matrix motivated by considerations of misspecification. In these models with parameterized stochastic discount factors, the sample moment conditions $g_N(\beta)$ can be interpreted as a vector of pricing errors associated with the parameter vector β. A feature of $W = V^{-1}$ is that, if the sample moment conditions (the sample counterpart to a vector pricing errors) happened to be the same for two models (two choices of β), the one for which the implied asymptotic covariance matrix is larger will have a smaller objective. Thus there is a *reward* for parameter choices that imply variability in the underlying central limit approximation. To avoid such a reward, it is also useful to compare models or parameter values in other ways. An alternative weighting matrix is constructed by minimizing the least squares distance between the parameterized stochastic discount factor and one among the family of discount factors that correctly price the assets. Equivalently, parameters or models are selected on the basis of the maximum pricing error among constant weighted portfolios with payoffs that have common magnitude (a unit second moment). See Hansen and Jagannathan (1997) and Hansen et al. (1995) for this and related approaches.

Selection Matrices

An alternative depiction is to introduce a *selection matrix* A that has dimension $k \times r$ and to solve the equation system:

$$A g_N(\beta) = 0$$

for some choice of β, which we denote b_N. The selection matrix A reduces the number of equations to be solved from r to k. Alternative selection matrices are associated with alternative GMM estimators. By relating estimators to their corresponding selection matrices, we have a convenient device for studying simultaneously an entire family of GMM estimators. Specifically, we explore the consequence of using alternative subsets of moment equations or more generally alternative linear combinations of the moment equation system. This approach builds on an approach of Sargan (1958, 1959) and is most useful for characterizing limiting distributions. The aim is to study simultaneously the behaviour of a family of estimators. When the matrix A is replaced by a consistent estimator, the asymptotic properties of the estimator are preserved. This option expands considerably the range of applicability, and, as we will see, is important for implementation.

Since alternative choices of A may give rise to alternative GMM estimators, index alternative estimators by the choice of A. In what follows, replacing A by a consistent estimator does not alter the limiting distribution. For instance, the first-order conditions from minimizing a quadratic form can be represented using a selection matrix that converges to a limiting matrix A. Let

$$D = E\left[\frac{\partial f(x_t, \beta_0)}{\partial \beta}\right].$$

Two results are central to the study of GMM estimators:

$$\sqrt{N}(b_N - \beta_0) \approx -(AD)^{-1}A\sqrt{N}g_N(\beta_0) \quad (5)$$

and

$$\frac{1}{\sqrt{N}}g_N(b_N) \approx \left[I - D(AD)^{-1}D\right]\sqrt{N}g_N(\beta_0). \quad (6)$$

Both approximation results are expressed in terms of $\sqrt{N}g_N(\beta_0)$, which obeys a central limit theorem, see (2). These approximation results are obtained by standard local methods. They require the square matrix AD to be nonsingular. Thus, for there to exist a valid selection matrix, D must have full column rank k. Notice from (6) that the sample moment conditions evaluated at bN have a degenerate distribution. Pre-multiplying by A makes the right-hand side zero. This is to be expected because linear combinations of the sample moment conditions are set to zero in estimation.

In addition to assess the accuracy of the estimator (approximation (5)) and to validate the moment conditions (approximation (6)), Newey and West (1987a) and Eichenbaum et al. (1988) show how to use these and related approximations to devise tests of parameter restrictions. (Their tests imitate the construction of the likelihood ratio, Lagrange multiplier, and the Wald tests familiar from likelihood inference methods.)

Next we derive a sharp lower bound on the asymptotic distribution of a family of GMM estimators indexed by the selection matrix A. For a given A, the asymptotic covariance matrix for a GMM estimator constructed using this selection is

$$\mathrm{cov}(A) = (AD)^{-1}AVA'(D'A')^{-1}.$$

A selection matrix in effect over-parameterizes a GMM estimator, as can be seen from this formula. Two such estimators with selection matrices of the form A and BA for a nonsingular matrix B imply

$$\mathrm{cov}(BA) = \mathrm{cov}(A)$$

because the same linear combinations of moment conditions are being used in estimation. Thus without loss of generality we may assume that $AD = I$. With this restriction we may imitate the proof of the famed Gauss–Markov theorem to show that

$$\left(D'V^{-1}D\right)^{-1} \leq \mathrm{cov}(A) \quad (7)$$

and that the lower bound on left is attained by any $\tilde{A}$ such that $\tilde{A} = BD'V^{-1}$ for some nonsingular B. The quadratic form version of a GMM estimator

typically satisfies this restriction when W_N is a consistent estimator of V^{-1}. This follows from the first-order conditions of the minimization problem.

To explore further the implications of this choice, factor the inverse covariance matrix V^{-1} as $V^{-1} = \Lambda'\Lambda$ and form $\Delta = \Lambda D$. Then

$$V^{-1}D\left(D'V^{-1}D\right)^{-1}D'V^{-1} = \Lambda'\left[\Delta(\Delta'\Delta)^{-1}\Delta'\right]\Lambda.$$

The matrices $\Delta(\Delta'\Delta)^{-1}\Delta'$ and $I - \Delta(\Delta'\Delta)^{-1}\Delta'$ are each idempotent and

$$\begin{bmatrix} \left[I - \Delta(\Delta'\Delta)^{-1}\Delta'\right] \\ \Delta(\Delta'\Delta)^{-1}\Delta' \end{bmatrix} \sqrt{N}\Lambda g_N(\beta_0)$$

$$\times \to \text{Normal}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix} I - \Delta(\Delta'\Delta)^{-1}\Delta' & 0 \\ 0 & \Delta(\Delta'\Delta)^{-1}\Delta' \end{bmatrix}\right).$$

The first coordinate block is an approximation for $\sqrt{N}\Lambda g_N(b_N)$ and the sum of the two coordinate blocks is $\sqrt{N}\Lambda g_N(\beta_0)$. Thus we may decompose the quadratic form

$$\begin{aligned} N[g_N(\beta_0)]'V^{-1}g_N(\beta_0) \quad &\approx N[g_N(b_N)]'V^{-1}g_N(b_N) \\ +N[g_N(\beta_0)]'V^{-1}D\left(D'V^{-1}D\right)^{-1}&D'V^{-1}g_N(\beta_0). \end{aligned} \tag{8}$$

where the two terms on the right-hand side are distributed as independent chi-square. The first has $r - k$ degrees of freedom and the second one has k degrees of freedom.

Implementation Using the Objective Function Curvature

While the formulas just produced can be used directly using consistent estimators of V and D in conjunction with the relevant normal distributions, looking directly at the curvature of the GMM objective function based on a quadratic form is also revealing. Approximations (5) and (6) give guidance on how to do this.

For a parameter vector β let $V_N(\beta)$ denote an estimator of the long-run covariance matrix. Given an initial consistent estimator b_N, suppose that $V_N(b_N)$ is a consistent estimator of V and

$$D_N = \frac{1}{N}\sum_{t=1}^{N}\frac{\partial f(x_t, b_N)}{\partial \beta}.$$

Then use of the selection $A_N = D_N'[V_N(b_N)]^{-1}$ attains the efficiency bound for GMM estimators. This is the so-called two-step approach to GMM estimation. Repeating this procedure, we obtain the so-called iterative estimator. (There is no general argument that repeated iteration will converge.) In the remainder of this section we focus on a third approach, resulting in what we call the continuous-updating estimator. This is obtained by solving

$$\min_{\beta \in \mathbb{P}} L_N(\beta)$$

where

$$L_N(\beta) = N[g_N(\beta)]'[V_N(\beta)]^{-1}g_N(\beta).$$

Let b_N denote the minimized value. Here the weighting matrix varies with β.

Consider three alternative methods of inference that look at the global properties of the GMM objective $L_N(\beta)$:

(a) $\{\beta \in \mathbb{P}: L_N(\beta) \le C\}$ where C is a critical value from a $X^2(r)$ distribution.
(b) $\{\beta \in \mathbb{P}: L_N(\beta) - L_N(b_N) \le C\}$ where C is a critical value from a $\chi^2(k)$ distribution.
(c) Choose a *prior* π. Mechanically, treat $-\frac{1}{2}L_N(\beta)$ as a log-likelihood and compute

$$\frac{\exp\left[-\frac{1}{2}L_N(\beta)\right]\pi(\beta)}{\int \exp\left[-\frac{1}{2}L_N(\beta)\right]\pi\left(\tilde{\beta}\right)d\tilde{\beta}}.$$

Method (a) is based on the left-hand side of (8). It was suggested and studied in Hansen et al. (1995) and Stock and Wright (2000). As emphasized by Stock and Wright, it avoids using a local

identification condition (a condition that the matrix D has full column rank). On the other hand, it combines evidence about the parameter as reflected by the curvature of the objective with overall evidence about the model. A misspecified model will be reflected as an empty confidence interval.

Method (b) is based on the second term on right-hand side of (8). By translating the objective function, evidence against the model is netted out. Of course it remains important to consider such evidence because parameter inference may be hard to interpret for a misspecified model. The advantage of (b) is that the degrees of freedom of the chi-square distribution are reduced from r to k. Extensions of this approach to accommodate nuisance parameters were used by Hansen and Singleton (1996) and Hansen et al. (1995). The decomposition on the right-hand side of (8) presumes that the parameter is identified locally in the sense that D has full column rank, guaranteeing that the $D'V^{-1}D$ is nonsingular. Kleibergen (2005) constructs an alternative decomposition based on a weaker notion of identification that can be used in making statistical inferences.

Method (c) was suggested by Chernozhukov and Hong (2003). It requires an integrability condition which will be satisfied by specifying a uniform distribution π over a compact parameter space. The resulting histograms can be sensitive to the choice of this set or more generally to the choice of π. All three methods explore the global shape of the objective function when making inferences. (The large sample justification remains local, however.)

Backing Off from Efficiency

In what follows we give two types of applications that are not based on efficient GMM estimation.

Calibration-Verification

An efficient GMM estimator selects the best linear combination among a set of moment restrictions. Implicitly a test of the over-identifying moment conditions examines whatever moment conditions are not used in estimation. This complicates the interpretation of the resulting outcome. Suppose instead there is one set of moment conditions for which we have more confidence and are willing to impose for the purposes and calibration or estimation. The remaining set of moment conditions are used for the purposes of verification or testing. The decision to use only a subset of the available moment conditions for purposes of estimation implies a corresponding loss in efficiency. See Christiano and Eichenbaum (1992) and Hansen and Heckman (1996) for a discussion of such methods for testing macroeconomic models.

To consider this estimation problem formally, partition the function f as:

$$f(x,\beta) = \begin{bmatrix} f^{[1]}(x,\beta) \\ f^{[2]}(x,\beta) \end{bmatrix}$$

where $f^{[1]}$ has r_1 coordinates and $f^{[2]}$ has $r^{-1}\ r_1$ coordinates. Suppose that $r_1 \geq k$ and that β is estimated using an A matrix of the form:

$$A = [A_1 \ \ 0],$$

and hence identification is based only on

$$A_1 E f^{[1]}(x_t,\beta) = 0.$$

This is the so-called calibration step. Let b_N be the resulting estimator.

To verify or test the model we check whether $g_N^{[2]}(b_N)$ is close to zero as predicted by the moment implication:

$$E f^{[2]}(x_t,\beta_0) = 0.$$

Partition the matrix D of expected partial derivatives as:

$$D = \begin{bmatrix} D_1 \\ D_2 \end{bmatrix}$$

where D_1 is r_1 by k and D_2 is $r^{-}\ r_1$ by k. Here we use limit approximation (6) to conclude that

$$\sqrt{N}g_N^{[2]}(b_N) \approx \left[-D_2(A_1D_1)^{-1}A_1I\right]\sqrt{N}g_N(\beta_0),$$

which has a limiting normal distribution. A chi-square test can be constructed by building a corresponding quadratic form of $r^{-1}\ r_1$ asymptotically independent standard normally distributed random variables. (When r_1 exceeds k it is possible to improve the asymptotic power by exploiting the long-run covariation between $f^{[2]}(x_t;\ \beta_o)$ and linear combination of $f^{[1]}(x_t;\ \beta_o)$ not used in estimation. This can be seen formally by introducing a new parameter $\gamma_o = E[f^{[2]}\ (x_t;\ \beta)]$ and using the GMM formulas for efficient estimation of β_o and γ_o.)

Sequential Estimation

Sequential estimation methods have a variety of econometric applications. For models of sample selection see Heckman (1976), and for related methods with generated regressors see Pagan (1984). For testing asset pricing models, see Cochrane (2001, Chaps. 12 and 13).

To formulate this problem in a GMM setting, partition the parameter vector as

$$\beta = \begin{bmatrix} \beta^{[1]} \\ \beta^{[2]} \end{bmatrix}$$

where $\beta^{[1]}$ has k_1 coordinates. Partition the function f as:

$$f(x,\beta) = \begin{bmatrix} f^{[1]}\left(x,\beta^{[1]}\right) \\ f^{[2]}(x,\beta) \end{bmatrix}$$

where $f^{[1]}$ has r_1 coordinates and $f^{[2]}$ has $r^-\ r_1$ coordinates. Notice that the first coordinate block only depends on the first component of the parameter vector. Thus the matrix d is block lower triangular:

$$D = \begin{bmatrix} D_{11} & 0 \\ D_{21} & D_{22} \end{bmatrix}$$

where

$$D_{ij} = E\left[\frac{\partial f^{[i]}(x_t,\beta_0)}{\partial \beta^{[j]}}\right].$$

A sequential estimation approach exploits the triangular structure of the moment conditions as we now describe. The parameter $\beta_0^{[1]}$ is estimable from the first partition of moment conditions. Given such an estimator, $b_N^{[1]}, \beta_0^{[2]}$ is estimable from the second partition of moment conditions. Estimation error in the first stage alters the accuracy of the second stage estimation, as I now illustrate.

Assume now that $r_1 \geq k_1$. Consider a selection matrix that is block diagonal:

$$A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}$$

where A_{11} has dimension k_1 by r_1 and A_{22} has dimension $k - k_1$ by $r^-\ r_1$. It is now possible to estimate $\beta_o^{[1]}$ using the equation system:

$$A_{11}g_N^{[1]}\left(\beta^{[1]}\right) = 0$$

or a method that is asymptotically equivalent to this. Let $b_N^{[1]}$ be the solution. This initial estimation may be done for simplicity or because these moment conditions are embraced with more confidence. Given this estimation of $\beta_o^{[1]}$, we seek an estimator $b_N^{[2]}$ of $\beta_0^{[2]}$ by solving:

$$A_{22}g_N^{[2]}\left[b_N^{[1]},\beta^{[2]}\right] = 0.$$

To proceed, we use this partitioning and apply (5) to obtain the limiting distribution for the estimator $b_N^{[2]}$. Straightforward matrix calculations yield,

$$\sqrt{N}\left(b_N^{[2]} - \beta_0^{[2]}\right) \approx -(A_{22}D_{22})^{-1}A_{22}\left[-D_{21}(A_{11}D_{11})^{-1}A_{11}I\right]\sqrt{N}g_N(\beta_0). \tag{9}$$

This formula captures explicitly the impact of the initial estimation of $\beta_0^{[1]}$ on the subsequent estimation of $\beta_0^{[2]}$. When D_{21} is zero an adjustment is unnecessary.

Consider next a (second-best) efficient choice of selection matrix A_{22}. Formula (9) looks just like formula (5) with A_{22} replacing A, D_{22} replacing

D and a particular linear combination of $g_N(\beta_0)$. The matrix used in this linear combination "corrects" for the estimation error associated with the use of an estimator $b_N^{[1]}$ instead of the unknown true value $\beta_0^{[1]}$. By imitating our previous construction of an asymptotically efficient estimator, we construct the (constrained) efficient choice of A_{22} given A_{11}:

$$A_{22} = B_{22}(D_{22})'\left(\left[-D_{21}(A_{11}D_{11})^{-1}A_{11}I\right]V\left[-\left[D_{21}(A_{11}D_{11})^{-1}A_{11}\right]'\right]\right)^{-1}$$

for some nonsingular matrix B_{22}. An efficient estimator can be implemented in the second stage by solving:

$$\min_{\beta^{[2]}} g_N^{[2]}\left(b_N^{[1]}, \beta^{[2]}\right)' W_N g_N^{[2]}\left(b_N^{[1]}, \beta^{[2]}\right)$$

for $V_N^{[2]}$ given by a consistent estimator of

$$V^{[2]} = \left(\left[-D_{21}(A_{11}D_{11})^{-1}A_{11}I\right]V\left[-\left[D_{21}(A_{11}D_{11})^{-1}A_{11}\right]'\right]\right)^{-1}$$

or by some other method that selects (at least asymptotically) the same set of moment conditions to use in estimation. Thus we have a method that adjusts for the initial estimation of $\beta^{[1]}$ while making efficient use of the moment conditions $Ef^{[2]}(x_t; \beta) = 0$.

As an aside, notice the following. Given an estimate $b_N^{[1]}$, the criterion-based methods of statistical inference described in section "Implementation Using the Objective Function Curvature" can be adapted to making inferences in this second stage in a straightforward manner.

Conditional Moment Restrictions

The bound (7) presumes a finite number of moment conditions and characterizes how to use these conditions efficiently. If we start from the conditional moment restriction:

$$E[f(x_{t+\ell}, \beta_0)|\mathcal{F}_t] = 0,$$

then in fact there are many moment conditions at our disposal. Functions of variables in the conditioning information set can be used to extend the number of moment conditions. By allowing for these conditions, we can improve upon the asymptotic efficiency bound for GMM estimation. Analogous conditional moment restrictions arise in cross-sectional settings.

For a characterizations and implementations appropriate for cross-sectional data, see Chamberlain (1986) and Newey (1993), and for characterizations and implementations in a time series settings see Hansen (1985, 1993), and West (2001). The characterizations are conceptually interesting but reliable implementation is more challenging. A related GMM estimation problem is posed and studied by Carrasco and Florens (2000) in which there is a pre-specified continuum of moment conditions that are available for estimation.

Conclusion

GMM methods of estimation and inference are adaptable to a wide array of problems in economics. They are complementary to maximum likelihood methods and their Bayesian counterparts. Their large sample properties are easy to characterize. While their computational simplicity is sometimes a virtue, perhaps their most compelling use is in the estimation of partially specified models or of misspecified dynamic models designed to match a limited array of empirical targets.

See Also

- Bayesian Methods in Macroeconometrics
- Rational Expectations Models, Estimation of
- Simulation-Based Estimation

Acknowledgments I greatly appreciate comments from Lionel Melin, Monika Piazzesi, Grace Tsiang, and Francisco Vazquez-Grande. This material is based upon work supported by the National Science Foundation under Award Number SES0519372.

Bibliography

Amemiya, T. 1974. The nonlinear two-stage least-squares estimator. *Journal of Econometrics* 2: 105–110.

Arellano, M. 2003. *Panel data econometrics*. New York: Oxford University Press.

Bontemps, C., and N. Meddahi. 2005. Testing normality: A GMM approach. *Journal of Econometrics* 124: 149–186.

Carrasco, M., and J.P. Florens. 2000. Generalization of GMM to a continuum of moment conditions. *Econometric Theory* 20: 797–834.

Chamberlain, G. 1986. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* 34: 305–334.

Chernozhukov, V., and H. Hong. 2003. An MCMC approach to classical estimation. *Journal of Econometrics* 115: 293–346.

Christiano, L.J., and M. Eichenbaum. 1992. Current real business cycle theories and aggregate labor market fluctuations. *American Economic Review* 82: 430–450.

Cochrane, J. 2001. *Asset pricing*. Princeton: Princeton University Press.

Cumby, R.E., J. Huizinga, and M. Obstfeld. 1983. Two-step two-stage least squares estimation in models with rational expectations. *Journal of Econometrics* 21: 333–335.

Eichenbaum, M.S., L.P. Hansen, and K.J. Singleton. 1988. A time series analysis of representation agent models of consumption and leisure choice under uncertainty. *Quarterly Journal of Economics* 103: 51–78.

Ghysels, E., and A. Hall. 2002. Editors' Introduction to *JBES* twentieth anniversary issue on generalized method of moments estimation. *Journal of Business and Economic Statistics* 20: 441.

Gordin, M.I. 1969. The central limit theorem for stationary processes. *Soviet Mathematics Doklady* 10: 1174–1176.

Hall, A.R. 2005. *Generalized method of moments*. New York: Oxford University Press.

Hall, P., and C.C. Heyde. 1980. *Martingale limit theory and its application*. Boston: Academic Press.

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50: 1029–1054.

Hansen, L.P. 1985. A method for calculating bound on asymptotic covariance matrices of generalized method of moments estimators. *Journal of Econometrics* 30: 203–238.

Hansen, L.P. 1993. Semiparametric efficiency bounds for linear time-series models. In *Models, methods and applications of econometrics: Essays in honor of A.R. Bergstrom*, ed. P.C.B. Phillips. Cambridge, MA: Blackwell.

Hansen, L.P. 2001. Method of moments. In *International encyclopedia of the social and behavior sciences*. New York: Elsevier.

Hansen, L.P., J. Heaton, and E. Luttmer. 1995. Econometric evaluation of asset pricing models. *Review of Financial Studies* 8: 237–274.

Hansen, L.P., and J.J. Heckman. 1996. The empirical foundations of calibration. *Journal of Economic Perspectives* 10 (1): 87–104.

Hansen, L.P., and R. Jagannathan. 1997. Assessing specification errors in stochastic discount factor models. *Journal of Finance* 52: 557–590.

Hansen, L.P., and K.J. Singleton. 1982. Generalized instrumental variables of nonlinear rational expectations models. *Econometrica* 50: 1269–1286.

Hansen, L.P., and K.J. Singleton. 1996. Efficient estimation of linear asset pricing models with moving average errors. *Journal of Business and Economic Statistics* 14: 53–68.

Hansen, L.P., Heaton, J.C., Lee, J. and Roussanov, N. 2007. Intertemporal substitution and risk aversion. In *Handbook of econometrics*, vol. 6A, ed. J. Heckman, and E. Leamer. Amsterdam: North-Holland.

Hayashi, F., and C. Sims. 1983. Nearly efficient estimation of time-series models with predetermined, but not exogenous, instruments. *Econometrica* 51: 783–798.

Heckman, J.J. 1976. The common structure of statistical methods of truncation, sample selection, and limited dependent variables and a simple estimator of such models. *Annals of Economic and Social Measurement* 5: 475–492.

Kleibergen, F. 2005. Testing parameters in GMM without assuming that they are identified. *Econometrica* 73: 1103–1123.

Newey, W. 1993. Efficient estimation of models with conditional moment restrictions. In *Handbook of statistics*, vol. 11, ed. G.S. Maddala, C.R. Rao, and H.D. Vinod. Amsterdam: North-Holland.

Newey, W. and McFadden, D. 1994. Large sample estimation and hypothesis testing. In *Handbook of econometrics*, vol. 4, ed. R. Engle, and D. McFadden. Amsterdam: North-Holland.

Newey, W.K., and K.D. West. 1987a. Hypothesis testing with efficient method of moments estimation. *International Economic Review* 28: 777–787.

Newey, W.K., and K.D. West. 1987b. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55: 703–708.

Ogaki, M. 1993. Generalized method of moments: econometric applications. In *Handbook of statistics*, vol. 11, ed. G.S. Maddala, C.R. Rao, and H.D. Vinod. Amsterdam: North-Holland.

Pagan, A.R. 1984. Econometric issues in the analysis of models with generated regressors. *International Economic Review* 25: 221–247.

Sargan, J.D. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26: 393–415.

Sargan, J.D. 1959. The estimation of relationships with autocorrelated residuals by the use of instrumental variables. *Journal of the Royal Statistical Society: Series B* 21: 91–105.

Singleton, K.J. 2006. *Empirical dynamic asset pricing: Model specification and econometric assessment*. Princeton: Princeton University Press.

Stock, J.H., and J.H. Wright. 2000. GMM with weak identification. *Econometrica* 68: 1055–1096.
West, K.D. 2001. On optimal instrumental variables estimation of stationary time series models. *International Economic Review* 42: 1043–1050.

Generational Accounting

Jagadeesh Gokhale

Abstract

Many government programmes transfer resources between different population groups. Programmes to provide retirement and health security levy taxes on workers to finance transfers to retirees. Initiating or expanding such programmes often redistributes wealth across generations by altering their lifetime tax burdens. Although standard budget measures such as national debt and deficits do not fully reflect them, such public intergenerational redistributions could substantially affect different generations' economic choices. Generational accounting measures the size of prospective net tax burdens facing different generations under current government tax and expenditure policies. It also analyses how those fiscal burdens would change under alternative policies.

Keywords

Aging populations; Budget deficits; Consumption; Fiscal burden; Fiscal policy; Generational accounting; Generational balance; Gifts; Government intertemporal budget constraint; Inheritance and bequests; Intergenerational transfers; Income taxes; Labour productivity; Labour supply; Labour-force participation; Lifetime net tax rates; National debt; Redistribution of income and wealth; Risk; Saving; Sensitivity analysis; Social insurance; Wealth

JEL Classifications

A1; H5; D4; D10; H23; H60; H69; E62

Before the 1990s, studies of the distributional impact of fiscal policies distinguished between groups according to their income, wealth or consumption at a point in time but not according to their life-cycle stage. Feldstein (1974) first pointed out the possibility of implementing large resource transfers across generations even under balanced government budgets. Nevertheless, notions about the impact of fiscal policies across generations remained limited to a presumed positive association between larger budget deficits and larger tax burdens on future generations.

Auerbach et al. (1991) developed generational accounting, a method for estimating the economic impact of fiscal policy on different cohorts – including future ones – distinguished by birth year and gender. With rapidly ageing populations in developed countries and growing costs of social insurance programmes that redistribute resources from younger to older generations, the demand for evaluating the intergenerational effects of government fiscal policies increased considerably. As a result, generational accounting is now used as a fiscal- analysis tool in dozens of countries.

Generational accounting (GA) is a method of estimating *prospective per capita lifetime net tax burdens* that different cohorts would face under existing fiscal policies. 'Prospective' means that fiscal burdens are evaluated over cohorts' remaining lifetimes; 'net tax' means that government transfers are subtracted from taxes; and 'lifetime' indicates that future dollar flows are actuarially discounted back to the present and aggregated into a summary measure of the fiscal burden in present value. Changes in the GAs of different cohorts arising from changes in government tax and spending policies measure fiscal policy-induced changes in those cohorts' lifetime resources.

Generational Accounting Method

Under current (year t) policies, the present discounted value of the government's projected purchases of goods and services (PVG_t) must be paid for out of the government's current net financial wealth (NW_t), the present value of net tax

payments by living generations (PVL_t), and the present value net tax payments by future-born cohorts (PVF_t). In this government intertemporal budget constraint,

$$PVG_t = NW_t + PVL_t + PVF_t, \qquad (1)$$

NW_t is calculated as the sum of past budget surpluses – which would be negative if past budgets mostly accrued deficits. The government's real assets, such as land, roads, buildings and public parks, are not included because that would require inclusion of a compensating term on the left-hand-side of Eq. (1) – the rental cost of the services those real assets provide.

For calculating PVL_t, official government projections of annual aggregate taxes and transfers are first distributed across officially projected populations using profiles of tax payments and transfer receipts by age and gender obtained from the latest available micro-data surveys. Per capita taxes and transfers for years beyond the government projection horizon are obtained by growing the terminal year's per capita values at the labour productivity growth rate underlying official aggregate projections.

Next, each living cohort's GA is calculated by actuarially discounting its projected net taxes per capita using cohort-specific mortality projections and an assumed rate of discount. Because fiscal dollar flows are more volatile than returns on government bonds but less volatile than private capital returns, an intermediate rate of interest is used. Multiplying each cohort's GA by its year-t population and aggregating across all cohorts yields PVL_t.

PVG_t is calculated by projecting government purchases of goods and services – such as administrative and judicial services, defence, and infrastructure – at current levels per capita using official population projections, and discounting those amounts back to year t. The term PVF_t in Eq. (1) is calculated as a residual.

Both PVL_t and PVG_t are calculated by projecting fiscal flows under unchanged policies. PVL_t equals the present value of net taxes that cohorts alive in year t would pay collectively if their fiscal treatment remained unchanged throughout their lifetimes. PVG_t indicates the size of the bill in present value for providing public goods and services at current levels for ever. To maintain the current fiscal treatment of living generations and current public goods and service levels for ever, the present value cost that future generations must pay equals $PVG_t - PVL_t - NW_t$.

Thus, generational accounting reveals the fiscal burden that future generations collectively face under current government fiscal policies. That burden does not necessarily equal the government's outstanding debt: $-NW_t$.

Estimating per capita fiscal burdens facing future-born generations requires knowing how it would be distributed among them. Generational accounting assumes, hypothetically, an equal distribution of the residual fiscal burden except for an adjustment for productivity growth. If we ignore gender differences for simplicity, the GA facing those born in year $t + 1$ is calculated as

$$GA_{t+1} = \frac{(PVG_t - PVL_t - NW_t)(1+r)}{\sum_{s=t+1}^{\infty} N_s\left[(1+g)/(1+r)\right]^{s-(t+1)}} \qquad (2)$$

Here, r represents the discount rate; g represents labour productivity growth; s represents future cohorts' birth years; and N_s represent their population sizes. In Eq. (2), the residual fiscal burden in present value as of period $t + 1$ is divided by the weighted sum of the population of future-born persons with weights based on r and g. The discount rate, r, is included in the weighting scheme to account for the differences in the timing of net tax payments by different future-born cohorts. Such weighting ensures that people born in period $s > t + 1$ pay lifetime net taxes that are $(1 + g)s - (t + 1)$ times larger than those paid by persons born in period $t + 1$.

Generational Accounts for the United States

By using projections from the Budget of the US government for fiscal year 2005 (with $t =$ fiscal

year 2004), applying a five per cent discount rate, and calculating US dollar amounts in constant 2004 dollars, PVG_t is estimated to be \$26.8 trillion; NW_t equals − \$4.4 trillion; and PVL_t equals 4.9 trillion. That leaves future generations to collectively pay \$26.3 trillion.

Table 1 shows GAs for selected US male and female cohorts with t = fiscal year 2004. They exhibit a standard life-cycle pattern: older cohorts face negative GAs – they receive benefits on net – and younger ones face positive GAs. Younger women have smaller GAs than men because of their lower labour-force participation and earnings. Very young cohorts with many years to go before paying taxes face considerably smaller GAs because of discounting. Older women receive larger net benefits in present value than older men despite their lower prior labour-force activity because they live longer and receive social insurance benefits based on their male spouses' earnings. The GA for those born in 2005 (year $t + 1$) equals \$333,200 per capita – considerably larger than that for 2004-newborns.

Lifetime Net Tax Rates and Generational Balance

Alternatively, fiscal burdens can be represented as *lifetime net tax rates* (LNTR) that different generations would face under the given assumptions. For future generations, $LNTRf = GA_s/PVE_s$, for all $s > t$, where PVE_s represents the present value as of period s of projected (pre-tax) labour earnings per capita for the cohort born in period s. Future labour earnings per capita are projected in a manner similar to that used for projecting taxes and transfers. Equation (2)'s distribution rule implies that both lifetime net taxes and lifetime earnings grow at the same rate for successive cohorts, implying that LNTR*f* applies to all future cohorts.

An important generational accounting concept is that of *generational balance.* It is derived by comparing the lifetime net tax rate facing year-t newborns, $LNTR_t = GA_t/PVE_t$, with LNTR*f.* Note that $LNTR_t$ is based on current tax and transfer policies extended throughout the lifetime of year-t newborns whereas LNTR*f* is a hypothetical rate imputed for future generations based on an equal growth- adjusted distribution of the residual fiscal burden across future-born cohorts. A finding of $LNTR_t < LNTRf$ would show current policy as being generationally out-ofbalance – one that levies a smaller LNTR on current newborns than would be required of future ones on average to balance the government's books. Thus, a policy that is generationally out of balance is also unsustainable.

Calculations based on the GAs shown in Table 1 reveal that US fiscal policy is considerably out of generational balance as of fiscal year 2004. The present value of lifetime earnings for males born in 2004 is estimated to be \$562,000, making $LNTR_{2004}$ equal to 18.5 per cent. For future-born cohorts, LNTR*f* equals 58.2 per cent. Continuing existing tax and spending laws for living generations would require future generations to bear fiscal burdens that are more than three times larger on average.

If current policy is out of generational balance (that is, if $LNTR_t < LNTRf$), GA machinery can also be used to calculate alternative policy changes that would restore generational balance. This exercise reveals the policy trade-offs involved in moving from a generationally out-of-balance policy to one that is balanced.

A large initial generational imbalance requires a large fiscal adjustment. Restoring generational

Generational Accounting, Table 1 Generational accounts for the United States (thousands of constant 2004 dollars)

Year of birth	Age in 2004	Male	Female
2005 (future-born)	−1	333.2	26.0
2004 (newborn)	0	104.3	8.1
1989	15	185.7	42.0
1974	30	201.3	30.2
1959	45	67.8	−54.1
1944	60	−162.6	−189.4
1929	75	−171.1	−184.1
1914	90	−65.0	−69.2

Source: Author's calculations based on data from Gokhale and Smetters (2006)

balance to US fiscal policy via income tax hikes would require average income tax rates to be 39 per cent larger. That is, federal income tax revenues that according to the US Congressional Budget Office (2006) amounted to 8.6 per cent of GDP in 2004 would have to be immediately and permanently increased to 11.9 per cent of GDP. Alternatively, federal discretionary outlays would have to be reduced immediately and permanently by 67 per cent.

Criticisms of Generational Accounting

Generational accounting has been subject to several criticisms. First, it measures the direct net costs of taxes and transfers but excludes the benefits derived from government public goods and service purchases. If the benefits from some purchases accrue much later, the average GA facing future generations may not accurately reflect their fiscal treatment under current policies. Second, generational accounting does not factor in the costs and benefits from government insurance provision.

These two criticisms indicate that generational accounting is not a 'utility measure' of the impact of fiscal policies on different generations. However, dynamic simulation studies suggest that changes in GAs correspond reasonably well to welfare gains and losses arising from policy changes.

Third, generational accounting ignores dynamic economic responses when estimating policy adjustments for restoring generational balance. However, its 'static' estimates constitute lower bounds of the required adjustments. For example, increasing income taxes would normally reduce labour supply and require a larger tax hike to achieve generational balance.

Fourth, to qualify as 'budget concepts' fiscal measures must show the implications of keeping policies unchanged. However, the generational balance measure employs a hypothetical policy for future generations. Gokhale and Smetters (2003) provide alternative fiscal and generational imbalance measures that do not involve hypothetical policies.

Fifth, generational accounting discounts future fiscal flows using a common discount rate whereas taxes and transfers may be subject to different degrees of policy and economic uncertainties. And sixth, it may be appropriate to use different discount rates for different cohorts because they face different risks. However, generational accounting studies include sensitivity analyses under alternative assumptions, including alternative discount rates.

Final Remarks

It is important to note that generational accounting tracks only the redistributive impact of government fiscal policies. It does not include the impact of private bequests and *inter vivos* gifts. In theory, private intergenerational transfers may substantially or fully offset government transfers. However, the weight of evidence, at least for the United States, suggests that such offsets are quite small.

A chief lesson from the generational accounting literature is that the frequently cited aggregate cash-flow measures of fiscal policy – such as the size of national debt and annual budget deficits – are uninformative and, indeed, may mislead policymakers about the true distributional and economic implications of current fiscal policies and policy changes.

To the extent that traditional deficit and debt measures miss significant policy- induced intergenerational redistributions – with potentially large effects on agents' economic choices such as consumption and labour supply – generational accounting calculations can provide useful information to policymakers and the public.

Generational accounting is also likely to prove useful in further economics and public-policy research. For example, generational accounts could be combined with other elements of wealth – human, non-human and private pension wealth – on a cohort basis to estimate whether changes over time in the cohort-distribution of resources are related to changes in cohort saving and labour force participation. Generational accounts could also be used to calculate changes in the degree of

cohort wealth annuitization for examining the extent of insurance against uncertain longevity.

In many countries, government programmes for providing insurance to the public against various types of economic risks are financially unsustainable. Uncertainty about prospective changes in taxes and transfers for correcting those fiscal imbalances constitute a major source of risk for households. Analyses using generational accounting may help in better understanding the extent to which government fiscal policies mitigate or exacerbate the economic risks facing different generations.

See Also

- ▶ Population Ageing
- ▶ Public Debt
- ▶ Public Finance

Bibliography

Auerbach, A., J. Gokhale, and L. Kotlikoff. 1991. Generational accounts: A meaningful alternative to deficit accounting. In *Tax policy and the economy*, vol. 5, ed. D. Bradford. Cambridge, MA: MIT Press/NBER.

Auerbach, A., J. Gokhale, and L. Kotlikoff. 1994. Generational accounting: A meaningful way to evaluate fiscal policy. *Journal of Economic Perspectives* 8(1): 73–94.

Congressional Budget Office. 2006. Historical budget data, 26 January, Table 4. Online. Available at http://www.cbo.gov/budget/historical.pdf. Accessed 28 June 2006.

Diamond, P. 1996. Generational accounting and generational balance: An assessment. *National Tax Journal* 49: 597–607.

Feldstein, M. 1974. Social Security, induced retirement, and aggregate capital accumulation. *Journal of Political Economy* 82: 905–926.

Gokhale, J., L. Kotlikoff, and J. Sabelhaus. 1996. Understanding the postwar decline in U.S. saving: A cohort analysis. *Brookings Papers on Economic Activity* 1996(1): 315–407.

Gokhale, J., and K. Smetters. 2003. *Fiscal and generational imbalances: New budget measures for new budget priorities*. Washington, DC: American Enterprise Institute Press.

Gokhale, J., and K. Smetters. 2006. Fiscal and generational imbalances: An update. In *Tax policy and the economy*, vol. 20, ed. J. Poterba. Cambridge, MA: MIT Press/NBER

Kotlikoff, L., and H. Fehr. 1996/97. Generational accounting in general equilibrium. *Finanzarchiv* 53: 1–27.

Kotlikoff, L. 1997. Reply to Cutler's and Diamond's views on generational accounting. *National Tax Journal* 50: 303–314.

Genovesi, Antonio (1712–1769)

Angelo Bertolini

Keywords

Genovesi, A.; Mercantilism; Population growth; Scarcity; Socialists of the chair; Value; Wealth

JEL Classifications

B31

Genovesi was born near Salerno and died at Naples: he took holy orders in 1736. In 1741 he taught metaphysics at the University of Naples. He was intimately acquainted with Bartolomeo Intieri, who induced him to follow Broggia and Galiani in the study of economics; and when, in 1754, by the advice of Intieri and with funds' liberally supplied by him, the teaching of economics, then termed mechanics and commerce, was established at Naples, Genovesi was called to the chair. He was 'the most distinguished and the most moderate of all Italian mercantilists. . . . Commerce was for him not an end only, but also a means by which the products of industry at large were brought to the right market. He, moreover, distinguished between useful commerce which exported manufactured goods and brought back in return raw material, and harmful commerce which exported raw material and imported foreign goods; he also insisted that useful commerce calls rather for liberty than for protection, while upon harmful commerce the strictest embargo should be laid, or at least it should as far as possible be bound hand and foor' (Cossa, *Introduction to Political Economy*, translation, p. 235).

These ideas, neither new nor original even in his time, were maintained by Genovesi in many of his works, and brought together, but without any systematic order, in his *Lezioni di Commercio ossia di Economia Civile* (Napoli, 1765, e. ii. ediz. 1768–70, 2 vols). Though the *Lezioni* do not form a regular treatise, they contain the author's opinions on the mercantilist system and the most important principles of economics, which he terms *Civile* '*la scienza che abbraccia le regole per rendere la sotto-posta nazione popolata, potente, saggia, polita*' (the science which embraces the laws which make a nation populous, powerful, wise, and cultured), limiting thus the science to the increase of population and the production of wealth.

As to population, Genovesi follows the mistaken principle of his times, exaggerating the advantage of a large population, proposing that government should encourage marriages by granting privileges and honours. He says that the population ought not only to be numerous but supplied with comforts, and he sees the relation between population and means of subsistence or production of wealth.

As a writer he is a mercantilist, though he does not regard money as the only form of riches; he says that the wealth of a nation is quite apart from the quantity of money treasured up.

He derives the idea of value from demand, distinguishing different degrees of demand according to their abstract importance in several categories, maintaining that a thing which satisfies a want repeatedly has a higher value than what satisfies only a few wants or the same only sometimes (*puo soddisfare ad un bisogno più volte, ha maggior prezzo che non quella, la quale o non puo soddisfare che pochi bisogni o al medesimo qualche volta*). What is able to satisfy a great want is of more value than what satisfies a small want (*una cosa fatta a soddisfare il maggior bisogno si apprezza più che quella la quale non é fatta che a soddisfare ad un minore*); and further he asserts that the quality of things influences the value. Graziani (*Storia della teoria del valore in Italia*, Milano, 1889, p. 108) justly remarks that in this Genovesi approaches the important question which Galiani answered: namely, why do luxuries generally cost more than necessaries? In this he is obliged to have recourse to the element of scarcity, a line of argument which he does not know how to reconcile with those previously mentioned. Genovesi's want of originality is obvious, as F. Ferrara has shown (*Bibl. dell' Econo.*,1*a*. S. vol. iii. Introduz.) in contradistinction to the exaggerated opinion which Bianchini held respecting him (*La scienza del ben vivere sociale*), since the Socialists of the Chair persist, erroneously, in considering him as a precursor of their opinions. This tendency is also attributed to Genovesi, as well as to Beccaria, Verri, and Romagnosi by the French socialist B. Malon; which is a further example of the errors of the socialists in their historical criticism of political economy.

See Also

► Mercantilism

Genuine Economic and Monetary Union

Iain Begg

Abstract

The severity of the euro crisis has already led to a series of reforms in economic governance, but it is accepted that further reforms are needed to make the euro more robust and resilient to asymmetric shocks. Proposals put forward by the EU's leaders to create a 'genuine' economic and monetary union, some of which have already been adopted, would deepen European integration to include a banking union, a greater degree of fiscal and political union and closer coordination of other economic policies. This article explains the background to the initiatives to establish a genuine economic and monetary union and assesses the progress towards achieving such a union.

Keywords

Economic and monetary union; Euro; European Union; Monetary integration; Optimum currency area; Sovereign debt crisis

JEL Classification

E6; E42; F15; F33; F36; H3; H6; O3; O31; O320; O380

If at first you don't succeed, try, try and try again – attributed to Robert the Bruce, King of Scotland, 1306–29.

In the beginning was a vision of monetary integration as an extension of European integration; then came the 'snake' and the European monetary system (EMS); next there was economic and monetary union (EMU) and the creation of the euro. Now, as they seek to rebuild after the tribulations of several years of economic and political crisis, Europe's leaders have a new approach: genuine economic and monetary union (GEMU). The first adjective in this expression invites (and has duly been accorded) mockery, for example from Willem Buiter who, in giving evidence on 18 June 2013 to Sub-Committee A of the House of Lords *European Union Committee* said in response to Q 57: 'I must say how poorly named GEMU is. Genuine economic and monetary union, as opposed to phoney or fake monetary union? Who dreamt this up? But never mind; we are stuck with it'.

However, it is clear that genuine economic and monetary union potentially constitutes a significant new departure for the economic governance of the euro area, with the potential to allow EMU to function sustainably. The sovereign debt crisis of 2010–12 had exposed many shortcomings in the structures and governance of the euro area and these had to be dealt with to ensure the future of the single currency. This article recalls the background and the initial responses to the euro crisis, then briefly summarises the measures adopted and planned to strengthen the governance and resilience of the euro. It then assesses their progress and the prospects for achieving an enduring euro.

Background to the Euro Crisis

Since the late 1960s, Europeans have been trying to find a way of establishing common monetary arrangements, starting with the blueprint provided by the committee chaired by Pierre Werner (Council and Commission, 1970) which set out plans for a single currency by 1980. Although agreement was reached, following the collapse of the Bretton Woods system, to set up a limited form of cooperation known as the 'snake in the tunnel' in 1972, it was short-lived, falling victim to the financial turmoil of the 1970s.

A fresh start was made in 1979 at the instigation of the leaders of France and Germany, Valéry Giscard d'Estaing and Helmut Schmidt, resulting in the much more comprehensive EMS, with its provisions for keeping exchange rates largely fixed between participating countries and intervention mechanisms aimed at countering the sorts of financial market pressures likely to upset an exchange rate arrangement. Although the EMS had to face a number of trials during the 1980s, Europe's leaders continued to push for more extensive monetary integration and eventually agreed – in the Maastricht Treaty concluded in 1990 – on the arrangements, which culminated at the end of the 1990s in full monetary union. Ironically, the agreement was rapidly followed in 1992 by the most severe crisis the EMS had faced, showing how damaging diverging economic trajectories can be for monetary integration.

The launch of EMU, initially with 11 participating countries, took place, as planned, from the late spring of 1998, when exchange rates were irrevocably fixed and the euro was created, with the deutschemark, lira, various francs, the guilder, the peseta and so on consigned to history. The newly created European Central Bank (ECB) became, almost overnight, one of the most important global economic actors. As EMU and the euro approached their tenth birthdays early in 2008, the consensus was that the single currency had been a remarkable achievement and was here to stay. Commemorative conferences and books abounded (for example, Buti et al., 2010), enabling Europeans to reflect on

achieving what many said was impossible or foolhardy. Although the sub-prime crisis had erupted in 2007 across the Atlantic, there was little sense that it posed a threat to the euro, and many in Europe took solace in what they saw as the more conservative approach of the European Central Bank and the financial supervisors in the EU, compared with their US counterpart, the Federal Reserve.

Then it all started to go wrong. First, many leading European banks discovered that they were not only much more exposed than they had realised to losses associated with the sub-prime crisis and the subsequent collapse of Lehman Brothers, but also that their national governments could not agree on how to rescue them, or (in some cases) afford to do so. The sudden 'discovery' late in 2009 that Greek public finances were in disarray, with projections of the annual deficit leaping from a manageable five or six percentage points of GDP to an unsustainable 15 points reignited a crisis that, by the summer of 2009, looked as though it was easing when leading countries such as Germany reported a return to growth.

The inability of the EU's leaders to resolve the Greek problem in a timely manner exposed systemic weaknesses in the governance of EMU, which then triggered contagion that engulfed Ireland and Portugal, and spread to threaten Italy and Spain before claiming Cyprus as its fourth victim. Thus, the initial financial crisis mutated into a sovereign debt crisis, engendering fears that the euro as a whole would unravel, with potentially cataclysmic consequences. In a succession of crisis meetings at which bailouts were agreed and massive new rescue funds were set up, Europe's leaders managed – just – to put out the fires and to ensure the survival of the euro. In this, they were greatly helped by the European Central Bank (ECB) and, not least, by the much quoted statement from its President, Mario Draghi (at the Global Investment Conference in London, 26 July 2012), that 'within our mandate, the ECB is ready to do whatever it takes to preserve the euro. And believe me, it will be enough'.

What Will It Take?

Despite the impression of procrastination conveyed to the rest of the world, far-reaching reforms were being undertaken by the EU to recast the governance of the euro. There were new legal provisions (commonly known as the 'six-pack') to enhance the oversight of Member States' fiscal policy and to instil greater fiscal discipline (including the *Stability and Growth Pact*), as well as to create a complementary *Macroeconomic Imbalances Procedure*, under which potential imbalances other than in the public finances would be monitored. In addition, a new permanent fund, the *European Stability Mechanism* (see http://www.esm.europa.eu/), was set up to provide financial assistance for crisis resolution, and various regulatory changes to improve financial stability were being implemented. However, a frequently heard critique has been that they are leading to an 'austerian' model of policy which will perpetuate economic weakness (Blyth 2013). Lane (2012) has identified a series of policy mistakes in the management of the euro debt crisis which have arguably made things worse.

A month prior to the Draghi speech, the European Council – the heads of state and government of the (then) 27 EU Member States – had agreed a draft of a document commonly known as the *Four President's Report* (European Council 2012a) which set out 'a vision for the future of the Economic and Monetary Union'. The report proposed four building blocks for a more robust framework for the governance of EMU in the pursuit of a 'genuine' EMU:

- An integrated financial framework to ensure a financially stable system
- An integrated budgetary framework with the dual aim of assuring fiscal discipline and developing new common fiscal policy instruments
- An integrated economic policy framework able to promote growth, employment and competitiveness in a manner consistent with the smooth functioning of EMU

- Enhancement of democratic legitimation and channels of accountability, justified particularly by the loss of national autonomy in budgetary and other economic matters as a direct consequence of greater top-down constraints on national autonomy in economic decision-making.

The first building block can be interpreted as *banking union*, the second as *fiscal union* and the last as at least a form of *political union*, while the third building block can be viewed as an elaboration of mechanisms to coordinate national economies more effectively and comprehensively. Proposals for developing each of these building blocks were subsequently presented in a 'blueprint' by the European Commission (2012) and taken further at the end of 2012 by the European Council (2012b). Both of the latter documents included proposals on the sequencing of the introduction of new measures, although it is noteworthy that they are by no means identical in what they propose. Some of the measures in the Commission blueprint would entail a significant extension of European integration and it is a moot point whether they are needed or, instead, are opportunistic bids for new powers by the EU level.

Why Was It Needed?

Despite the flurry of reforms in 2011 and 2012, the impetus for going much further came from an understanding that there had to be a more fundamental rethinking of economic and monetary union as a defining element of the European integration 'project'. There have been extensive efforts to understand the origins of the crisis (for instance, Eichengreen 2012; De Grauwe 2013) and to assess whether the policy responses were well conceived, timely and sufficiently comprehensive. Answers have been cautious, with many (especially from across the Atlantic) sceptical about the euro's long-term viability, and some even suggesting that it would be better for the euro to collapse quickly rather than endure a lingering demise. A comprehensive retort is offered by Eichengreen (2007, 2014), who sketches out the potentially calamitous outcomes of euro breakdown (or even of exit by an existing member) in terms of banking crisis, capital flight, inflationary pressures and the need for exchange controls. His argument is that although nothing is forever, euro membership comes close.

Nevertheless, there is a consensus that the original architecture of EMU had to change. Four categories of flaws in what might be called 'EMU1' can be distinguished. The first can loosely be summed up in the title of an article by Paul Krugman (2012): 'Revenge of the optimum currency area'. Europe was generally acknowledged not to be an optimum currency area (OCA), and there was reluctance to put in place measures that might have mitigated the effects of this lack, notably fiscal mechanisms that might have attenuated asymmetric shocks. A similar metaphor is used in the title of a paper by Barry Eichengreen (2014): 'the theory of optimum currency areas bites back'. However, his analysis dwells more on the shortcomings of the theoretical insights themselves, now over 50 years old, than on how EMU neglected these insights. Indeed, Eichengreen identifies crucial ways in which the experience of EMU reveals problems of monetary integration simply not anticipated in OCA reasoning.

According to what McKinnon (2004) has called the 'Mundell I' version of OCA theory, when a shock hits an economy that is part of a monetary union, it can be offset in a number of ways, notably through factor mobility and automatic stabilisation brought about by fluctuations in net fiscal transfers from the rest of the union. A downturn will reduce tax flows to the central government, while public expenditure programmes mediated at the highest level of government disburse more in the affected region because of entitlements assured by the highest level of government (for example unemployment benefits). The trouble for the EU is that there is no such fiscal instrument, because the only budget at EU level is small (at just 1% of GDP), inflexible and required to balance, preventing it from being used to provide a fiscal stimulus. Labour mobility

is notoriously low in the EU, and such mobility as there is probably even has a perverse effect (as pointed out by Eichengreen 2014) in so far as it is qualified and skilled workers who leave in a downturn, denuding the economy in difficulty of growth potential. If, in addition, the distressed economy is unable to use its own fiscal policy to stimulate recovery (as proved to be the case in the euro crisis), the difficulties are compounded.

The second area of concern is the effects of EMU on the supply side of the economy. In the absence of the exchange rate, economic adjustment has to occur through other policy channels. With fiscal policy constrained by either the rules of EMU or by market reluctance to lend to countries experiencing a downturn, autonomy in fiscal policy has proved to be relatively limited. This meant that adjustment had to come from structural policies. Indeed, there was an expectation that EMU would oblige governments to pay greater attention to such policies, but the reality was that the sizeable fall in interest rates enjoyed prior to the crisis by countries such as Italy meant that they had windfall gains for their public finances. As a result, reform pressures abated and the resilience and flexibility of the economy was eroded. It can be argued that the imperative of structural reform today is all the greater because this phenomenon was underestimated.

The easy macroeconomic conditions and the absence of market pressures in the middle of the euro's first decade also led to a false sense that all was well. On the contrary, as many subsequent analyses have shown, divergence was occurring, particularly in unit labour costs, rendering many of the countries that have since struggled less competitive. By 2008, very large external imbalances had arisen inside the euro area, the most extreme being the 15% of GDP current account deficit in Greece in 2009. These imbalances led to conflicting views on who was to blame: the profligate periphery or the mercantilist core? For Germany, especially given its economic weight and export-orientated economic model, there are many dilemmas to confront around whether it shares responsibility for the crisis or can blame it on bad policy elsewhere (Bonatti and Fracasso 2013).

Financial integration in Europe is the source of a third set of flaws and, moreover, gave rise to problems that were much less predictable (Pisani-Ferry 2012a). The 'Mundell II' version of OCA posits a different channel of adjustment, operating through financial markets. In this variant, insurance or changes in net savings can come to the rescue by providing financial flows which cushion the effects of a downturn. The EU has assiduously promoted financial integration and has had considerable success in some respects, for instance in fostering cross-border bond markets, However, the holdings of sovereign debt by banks in creditor countries became an obstacle rather than a help when market sentiment turned against countries facing severe adjustment pressures. Cross-border flows at the retail level, by contrast, were negligible and thus could not act as they do in more integrated monetary unions, such as the US or Canada, to mitigate a downturn. It is important to note that the inability to deal with shocks is a problem not just for the country affected, but also because of the rapid spillover of fiscal stress to other countries – described by Allard et al. (2013) as a systemic problem.

Much has been written about the 'doom loop' between banks and sovereigns in the euro area, in which problems in the banking sector lead to public intervention, causing a deterioration of public finances, while bank holdings of suspect sovereign debt have weakened banks' capital base. The extent of financial contagion from one vulnerable member state to another was also an unexpected phenomenon, yet had become the most pressing problem by the end of 2011, worsening as markets piled the pressure on Italy and Spain, causing the spreads on bonds denominated in euros issued by these countries to rise dramatically, relative to Germany (Favero and Missale 2012). Private debt was also part of the story in some member states, such as Ireland and Spain, pointing to macro-prudential failings. As Constancio (2014: 251) observes, the banking crisis and the sovereign debt crisis 'create a story of two "debt overhangs"', a position he contrasts with what he suggests has been an incorrect, but prevailing, narrative of blaming fiscal indiscipline (mainly) in peripheral countries. Moreover, as

pointed out by Pisani-Ferry (2012b), there is a new 'impossible trinity' because of the conjunction in the original design of EMU of:

- The prohibition on direct monetary financing of the debts of member states which appears to preclude the ECB from direct purchases of sovereign debt.
- The fact that there is no collective responsibility for public debt, such that member states in difficulty are susceptible to market pressures much more rapidly than if there were a common borrowing capability. Some form of eurobond, jointly and severally guaranteed by all member states (at least of the euro area), is the eventual answer.
- The interdependence between sovereigns and banks in each member state can result in banks becoming fragile if they hold their country's public debt, while the fragility of the banks undermines the borrowing status of the sovereign that has to stand behind them. Sovereign bonds tend to be thought of as safe assets, but the problems in Ireland and Spain have shown that market sentiment can turn quickly, leading to a vicious circle, especially in smaller countries.

In the same way that monetary union offered a way out of the original Mundell impossible trinity, an underlying rationale for banking union, especially, but also elements of the other facets of GEMU, is to provide part of the solution to the Pisani-Ferry one.

The fourth major flaw was one of governance. EMU had rules, but they were not respected, partly because of how institutional boundaries functioned, but partly also because the rules themselves were open to criticism. More generally, the EU has been beset by political economy obstacles to comprehensive solutions (Begg 2013) and there has been a lack of understanding of political and institutional differences or difficulties among the Member States and their impact on preferences (Bordo et al., 2013). German leaders, for instance, have had to be sensitive to possible complaints to the Federal Constitutional Court (of which there have been many) being upheld in a way that would have derailed proposed solutions.

The crisis also exposed the inability of the euro area members to react quickly, perhaps because a lack of experience in handling so acute a crisis hampered an effective response. Manifestly, a crisis of this magnitude was not anticipated by the authors of the Treaty, nor does it seem to have been foreseen in the design of the specific instruments and institutional arrangements that make up the economic governance frameworks (Lane 2012). Crisis management and resolution mechanisms were also conspicuously missing. In short, there was no readily available toolkit. Some of the GEMU proposals hint at the establishment of an enhanced ability to manage crises. But proposals for a European Monetary Fund (for example, Gros and Mayer 2010; Schulmeister 2013) though apparently endorsed by German Finance Minister Schäuble (the *Financial Times*: 'Why Europe monetary union faces biggest crisis' 11 March 2010), have met little enthusiasm.

Progress on Achieving GEMU

The Commission blueprint offers an extensive list of specific proposals, but a number of these are speculative rather than firm. The Commission also distinguishes between proposals that can be implemented quickly, over the medium term, and over the longer term. Not surprisingly, many of the proposals have attracted dissent, and it is open to question whether they will be adopted, substantially amended or watered down, or simply abandoned. The resulting uncertainty inevitably complicates any assessment of what their implications will be for the final shape of GEMU. There is also ambiguity in the usage of many of the key terms around GEMU, worthy of Humpty Dumpty's assertion in Lewis Carroll's book *Through the Looking Glass*: 'When I use a word, it means just what I choose it to mean, neither more nor less'.

Banking union has made the most progress, but it has proved much more difficult to agree fiscal union, partly because member states have very divergent views on what is needed and (often more tellingly) who should pay or shoulder the

risk. Proposals for closer coordination have exposed deep cleavages among member states and between the member states and the EU institutions, while political union is barely off the drawing board.

Banking Union

A banking union would, if most of its advocates had their way, consist of three main elements which were absent in the original EMU framework: an integrated approach to bank supervision; a common means of resolving failing banks; and the establishment of a common system of deposit insurance. Numerous commentators (see, for example: Pisani-Ferry and Wolff 2012; Obstfeld 2013; Leblond 2014) have highlighted the sheer scale of banking in relation to the fiscal resources of certain member states, as well as the extent of cross-border financial connections, as features incompatible with national oversight. In the absence of banking union, it fell to the ECB to underpin financial stability, despite the lack of a formal lender of last resort function, and part of the logic of banking union was to redress an overreliance on the central bank. As Muellbauer (2013) stresses, expecting a central bank to fulfil this role indefinitely is not defensible.

After typically complicated negotiations and the careful crafting of compromises, two elements of banking union have now been settled. First, the creation of the single supervisory mechanism (SSM) in 2013 conferred significant new powers on the ECB as the pinnacle of a federal-ish system of oversight of banks, but at German insistence, only the largest and systemically most important banks are to be directly supervised by the ECB. New 'bail-in' rules determining the order in which different classes of creditor would lose their money were agreed and complement the deal reached at the end of 2013 on a single resolution mechanism (SRM). The latter includes provision for a new single resolution fund which will be built up from levies on the banking sector and is intended to reduce the probability of future calls on taxpayers to rescue banks.

Although the single supervisory mechanism and the single resolution mechanism constitute significant steps towards banking union, they do not immediately resolve the doom loop because there is, as yet, no consensus on establishing a supranational fiscal backstop. Without one, as Obstfeld (2013) shows, the whole structure will be vulnerable, but for the time being, only national backstops are available to use in this way. In addition, the single resolution fund will only build up slowly (over a decade) to reach its target scale of €55 billion, a level that many critics consider to be far too low to deal with potential problems in what remains a fragile banking system subject to considerable risks in the short-term.

Common deposit insurance remains a distant prospect, having essentially been vetoed by creditor nations in the EU because of worries about moral hazard. The likely knock-on effect is that the euro area will remain less integrated financially (Howarth and Quaglia 2013), restricting the scope for private financial flows to accommodate asymmetric shocks. Indeed, the trend since the crisis has been for a *de facto* growth of barriers to financial integration as credit institutions retrench within national boundaries. This has had the further knock-on effect of restricting the supply of credit in weaker economies, inhibiting their capacity to restore growth and employment.

Fiscal Union

There are many possible definitions of fiscal union in the context of EMU (Fuest and Peichl 2012; Tommaso Padoa-Schioppa Group 2012; Allard et al., 2013). At one extreme it could mean little more than an intensification of the rules already in place to discipline deficits and public debt, as amended by the 'six-pack', referred to above, and subsequent measures. At the other, it could encompass substantial flows between the supranational level and the constituent parts of the EMU, in other words a transfer union in a form that most Germans just about tolerate domestically, but are adamantly opposed to for the EU on the not unreasonable grounds that they would generally have to foot the bill. A more limited fiscal union could entail some mutualisation of debt and the creation of funds to alleviate fiscal stress in a country subject to temporary difficulties.

Adding to the complexity of the issues is a lack of clarity about the underlying purpose of a supranational fiscal capacity. The current EU budget supports public investment, but is regarded by many as mainly having a redistributive function, and manifestly has no role in macroeconomic stabilisation akin to the federal level elsewhere (Begg 2009). An additional fiscal capacity for stabilisation purposes was envisaged in early plans for EMU, including the MacDougall Report (1977), and discussed again in *One Market, One Money*, a wide-ranging study by the European Commission (1990) which examined various aspects of how the EMU agreed at Maastricht should develop. A fiscal capacity has, however, unfailingly fallen foul of objections from net contributors to the EU budget to any increase. For GEMU to succeed in overturning these objections will not be easy.

It is nevertheless firmly advocated – albeit without much sense of what precise form it might take – in the Commission blueprint for GEMU. The proposals imply a stabilisation function confined to assisting individual member states in mitigating the effects of an asymmetric shock, but not the stabilisation of the eurozone economy as a whole. The attraction of a limited stabilisation instrument is that it requires fewer resources, but it would not provide the scope for a collective Keynesian response to any future recession. Latterly, there has been renewed interest in a risk-sharing policy mechanism that would be triggered by variations in unemployment rates from an appropriate benchmark (Andor et al., 2014; Caudal et al., 2013). Its proponents argue that it would not result in persistent net transfers, but critics are sceptical.

There has been extensive discussion of the need for common debt issuance by euro area members (in effect a euro version of the US Treasury Bond), whether called eurobonds (a term now so toxic in Germany as to be unacceptable), Stability Bonds (European Commission 2011) or something else. The key feature of such a scheme would be joint and several guaranteeing of the debt, meaning in practice that net creditors would bear some of the risk of net debtors. Some form of debt mutualisation is included in the Commission blueprint for GEMU, but – reflecting the ambivalence around the concept – as one of the medium to longer term objectives for the fiscal union dimension. It is a broadening of EMU which many commentators, drawing on the US experience and the clear benefits associated with a strong and stable bond for the cost of financing public debt, regard as essential. Opponents fret about the moral hazard problems likely to arise if states which already borrow too readily have incentives to borrow still more.

A number of proposals have attempted to reconcile calls for a mutualised debt instrument and worries about moral hazard. Solutions include capping the share of debt as a proportion of GDP, as in the blue bond/red bond scheme of Delpla and von Weizsäcker (2011), or the conditional bonds proposed by Muellbauer (2013) which would carry a common interest coupon, but require the more risky member states to pay an insurance premium to insure (and politically placate) the other underwriting countries. Others have suggested that the best way forward is, instead, to lower the legacy debts of heavily indebted countries through some form of managed default. Examples are the debt redemption pact proposed by the German Council of Economic Experts (2012), allowing a 'once-only' debt reduction, or the 'politically acceptable debt restructuring' (PADRE) idea from Pâris and Wyplosz (2014), under which an agency (the ECB or some other entity nevertheless relying on the ECB) buys up the excessive debt.

In both these examples, the proponents suggest ways to forestall moral hazard. In all cases, some form of institutional development involving the creation of a euro area debt agency or a fully-fledged Treasury will be required – never easy to settle rapidly in the EU. It is likely to require political oversight from a more powerful EU or euro area finance minister equivalent, especially if revenue has to be raised as a backstop to borrowing operations.

Closer Coordination

The rationale for closer coordination is to ensure that negative spillovers from inappropriate supply-side reforms (or an absence of reforms)

are avoided. The intention is to improve the resilience of a country inside the EMU to asymmetric shocks and their knock-on effects on others. If, as the Commission argues happened during the euro crisis, these knock-on effects have a substantial negative impact, they can become a threat to the EMU as a whole. To avoid such an outcome, the aim of *ex ante* coordination is to identify potential problems and to resolve them before the reform is implemented. In a proposal published in March 2013, the Commission set out a list of the policy areas to be covered, encompassing 'product, services and labour market reforms as well as certain tax reforms' (European Commission 2013a: 3), and also refers to financial markets as a source of spillover or contagion. The communication stresses, separately, the advantages of learning from each other.

A 'Convergence and competitiveness instrument', proposed as one of the developments of the third leg of GEMU, was put forward in a Commission communication released in March 2013, but has proved to be controversial. The original idea was to make additional resources available 'to support Member States in the sometimes difficult process of implementing structural reforms, provided that the recipient country accepted a reform programme agreed with the Commission (European Commission 2013b: 6). The Commission proposal stated that the support would be limited to reforms that might affect other EMU participants or the euro area as a whole. It would be a contract in the sense that adherence to the programme would be the condition for receiving the financial support – in effect a form of conditionality, as has routinely been practised over many years by the IMF. The idea was poorly supported at the December 2013 European Council and was postponed for a year. An indication of some of the sensitivities is visible in the apparent decision to rename the instrument as 'Partnerships for Growth, Jobs and Competitiveness' and to insist that these will be about increasing 'the level of commitment, ownership and implementation of economic policies and reforms in the euro area Member States'.

Political Union

A political union to underpin EMU has long been advocated (De Grauwe 2006, 2013), but has faced the inevitable difficulty that it is hard to agree what it means. An irony is that both advocates and critics of monetary union have identified political union as a dimension of EMU needed for it to function more effectively – albeit drawing opposing conclusions about whether a move in a federal direction is appealing. However, there is a lack of consensus – visible, for example, in conflicting French and German positions – about what comes first: political or fiscal union. There has been a proliferation of new rules and institutions, which are expected to transform economic governance, but there is as yet not a clear picture of how they will be politically controlled (Belke 2013). The more intense oversight of member states as well as the power to impose financial sanctions on delinquents are two outcomes of governance reforms that alter the balance of power in the EU. They will cause concerns about legitimacy and accountability. Add in calls for more comprehensive new mechanisms for sharing risk (Vallée 2014) and the growing social pressures on governments, and it becomes clear that there are difficult challenges around the political union dimension of GEMU.

Other than fairly general statements about the importance of democratic legitimation, only vague proposals for an increased European Parliament (EP) role are made in the Commission blueprint for GEMU. The rationale for seeking an EP role is that because many of the new powers are exercised at EU (or euro area) level, there is an equivalence logic, as understood in normative fiscal federalism approaches, in having legitimation at the same level of governance. As the Commission (2012: 35) states: 'a further strengthened role of EU institutions will therefore have to be accompanied with a commensurate involvement of the European Parliament in the EU procedures'. Both the Four Presidents' report and the Commission blueprint for genuine economic and monetary union nevertheless allude to the need for national parliaments to ensure the legitimacy of Member actions, but suggest no greater role.

One of the complications around the moves towards GEMU is how to accommodate non-members among the rest of the EU28. Two of them (Denmark and the UK) have a formal opt-out from the Treaty obligation to join EMU, but all the rest are, in principle 'in derogation', which means that they are expected to join the single currency in due course. While the years of crisis have seen many of them postpone any moves in this direction, the number of euro area members continues to grow, with Lithuania now set to join in January 2015. However, there are other dividing lines. Only the Czech Republic and the UK did not sign the separate *Treaty on Stability Coordination and Governance* (TSCG) ratified in 2012. This treaty provides a legal basis for some of the GEMU measures.

Although the UK has supported the implicit deepening of European economic governance needed to make the euro stable, it has stated emphatically that it will not participate in banking union or fiscal union (see evidence summarised in House of Lords 2014). Other member states, by contrast, have agreed to participate in banking union, including some, such as Sweden or Poland, which have often been sympathetic to UK positions on European matters.

Can 'Genuine' Render EMU Safe(r)?

As is so often the case in the EU, many of the measures in the Commission (2012) blueprint are being adopted at a slower pace and less comprehensively than envisaged. As always, too, it is battles over the sharing of burdens and risks that lie behind the procrastination. Nevertheless, a more robust and resilient EMU is being constructed; it may be less than optimal, but the threat of unravelling seems to have receded. As Benjamin Cohen (2012: 689) bluntly puts it: 'the euro will neither fail nor succeed. Defective but defended, it will simply endure'. Mongelli (2013) offers a more sanguine view of the euro as an underpinning for resolving what he sees as a crisis that had its roots elsewhere, notably in flawed governance.

That said, economic divergence is and will remain an awkward problem from two perspectives. First, it will make the task of curbing the imbalances that lie behind the crisis that much harder, and if – as seems to be happening – the divergence accentuates virtuous and vicious cycles, the challenge will be all the greater. Second, it greatly complicates one-size-fits-all policies such as monetary policy, because different economies will require much more tailored policy mixes. Aggravating factors are that country risk has widened the interest rate spread between troubled and stable economies, deterring investment in the South and stimulating it in Germany and other northern member states, while the debt overhang of many euro members is large. This is not a recipe for convergence or common policies.

For the foreseeable future, what has been, or is likely to be, agreed on banking union and fiscal union will not be enough to resolve completely some of the known shortcomings in EMU. A more open question is whether they are sufficient to enable the euro to function tolerably successfully. Thus far, the dire predictions of the Cassandras have not come true. The euro has not collapsed, nor has any member exited it, despite the chorus of voices urging countries like Greece to do so. The main reason is that there continues to be a strong political commitment to the euro, although no one should be in any doubt that a disorderly unravelling of the euro would have severe consequences.

What are largely missing from the GEMU blueprint, however, are proposals on reform or recasting of the role of the ECB. The ready explanation is that the strong independence conferred by the Treaty on the ECB means that the Commission and the European Council tread carefully for reasons of constitutional propriety where monetary matters are involved. The ECB is a powerful economic actor, reflecting its high degree of independence, yet its ability to act as a lender of last resort or to emulate the sorts of unconventional monetary policies pursued by its counterparts in the USA, Japan or the UK has been circumscribed. Although the ECB often acted at

crucial times to mitigate the crisis when governments were slow to act, its action mainly bought time (Bini-Smaghi 2014), but it is unrealistic to expect the central bank to solve the underlying problems that caused the crisis. This is one area where GEMU is in a bind, because it does not make clear what role is now expected of the monetary authorities. Indeed, Bini-Smaghi expresses the concern that, because ECB action has defused the crisis, governments may be less willing to take hard choices and thus risk triggering a fresh round of crisis. As Honkapohja (2014: 265) observes 'a basic lesson from the Nordic and other financial crises is that once the crisis has erupted, only hard choices for economic policy remain'.

However, there have been moves by, or involving, the ECB which nevertheless address some of the shortcomings in EMU and can thus be considered part of the spirit of GEMU. It has acquired additional responsibilities associated with prudential supervision at both the macro and micro levels (which could be interpreted as a *functional broadening*) but also a *deepening* in terms of having a generally more central and profound role in the overall conduct of economic policy, including taking part in the 'Troika' missions overseeing adjustment in countries accorded a bailout. In this regard, as contributions to the discussion in Chapter 5 of Blinder et al. (2013) highlight, the ECB has moved closer to the pivotal governance roles of the Federal Reserve and the Bank of England: that is, it has become a more 'normal' central bank.

A Last Word

GEMU, if substantially realised, would strengthen monetary union and, together with continuing 'whatever it takes' political commitments, ensure the sustainability of the euro. However, it is by no means a complete solution and there are too many loose ends and compromises. This is not because there is a dearth of analysis or understanding of what might be needed, but the trouble with many of the criticisms of the architecture of the euro is that the remedies proposed, as one such critic conceded, 'are very far removed from the present policy positions, and remote from what could be viewed as politically feasible' (Sawyer 2013: 11). This is the nub of the political economy problem confronting the euro: there may be an ideal setup for a monetary union, but it is not one that the all the members of the euro area are ready to accept, or at least not yet.

The adjective 'genuine' is a tricky one given its connotations with avoiding fakes, being truly of value and being legitimate, as well as implying sincerity. As applied to EMU, it is above all about whether the extensive governance changes introduced since the euro crisis first surfaced late in 2009 have done enough to correct the flaws in the original design of the single currency. A cautious verdict is that they have and thus that the outcome will be a 'durable' EMU – an expression (coined by Lord Vallance, a member of sub-Committee A of the House of Lords European Communities Committee) that may be more suitable than 'genuine'. Until the bad news is conclusively out of the European banking system, though, and the economies of the euro area return to steady growth, vulnerabilities will remain.

See Also

- Bretton Woods system
- European Monetary Integration
- European Monetary Union
- Stability and Growth Pact of the European Union

Bibliography

Allard, C., P. Koeva Brooks, J.C. Bluedorn, F. Bornhorst, K. Christopherson, F. Ohnsorge, and T. Poghosyan. 2013. Towards a fiscal union for the euro area. *IMF Staff Discussion Note 13/09*. https://www.imf.org/external/pubs/ft/sdn/2013/sdn1309.pdf.

Andor, L., S. Dullien, X. Jara, H. Sutherland, and D. Gros. 2014. Forum: Designing a European unemployment insurance scheme. *Intereconomics* 49(4): 184–203.

Begg, I. 2009. *Fiscal federalism, subsidiarity and the EU budget review*. Policy report, 2009, 1. Stockholm: Swedish Institute for European Policy Studies.

Begg, I. 2013. *The political economy of sovereign debt crises*. SUERF studies, 2013/01: 7–13.

Belke, A. 2013. Towards a genuine economic and monetary union – Comments on a roadmap. *Politics and Governance* 1: 48–65.

Bini-Smaghi, L. 2014. Austerity: A threat to democracy. *International Spectator* 49(1): 7–17.
Blinder, A., T.J. Jordan, D. Kohn, and F. Mishkin. 2013. *Exit strategy. Geneva report on the world economy, 15*. London: CEPR.
Blyth, M. 2013. *Austerity: The history of a dangerous idea*. Oxford: Oxford University Press.
Bonatti, L., and A. Fracasso. 2013. The German model and the European crisis. *Journal of Common Market Studies* 51(6): 1023–1039.
Bordo, M.D., L. Jonung, and A. Markiewicz. 2013. A fiscal union for the euro: Some lessons from history. *CES-ifo Economic Studies* 59(3): 449–488.
Buti, M., S. Deroose, V. Gaspar, and J. Nogueira Martins. 2010. *The euro: The first decade*. Cambridge: Cambridge University Press.
Caudal, N., N. Georges, V. Grossmann-Wirth, J. Guillaume, T. Lellouch, and A. Sode. 2013. *Un budget pour la zone euro*. Trésor Éco Lettre No. 120.
Cohen, B.J. 2012. The future of the euro: Let's get real. *Review of International Political Economy* 19(4): 689–700.
Constancio, V. 2014. The European crisis and the role of the financial system. *Journal of Macroeconomics* 39(part B): 250–259.
Council and Commission of the EC. 1970. Report to the Council and the Commission on the Realisation by Stages of Economic and Monetary Union in the Community – The Werner Report. *Bulletin of the EC*, 7–1970, Suppl.
De Grauwe, P. 2006. What have we learnt about monetary integration since the Maastricht treaty? *Journal of Common Market Studies* 44(4): 711–730.
De Grauwe, P. 2013. *Design failures in the Eurozone: Can they be fixed*? LSE 'Europe in Question' discussion paper series no. 57/2013.
Delpla, J., and J. von Weizsäcker. 2011. *Eurobonds: the blue bond concept and its implications*. Bruegel policy contribution 2011/02.
Eichengreen, B. 2007. *The break-up of the euro area*. NBER working paper no. 13393.
Eichengreen, B. 2012. European monetary integration with benefit of hindsight. *Journal of Common Market Studies* 50(S1): 123–136.
Eichengreen, B. 2014. *The Eurozone crisis: The theory of optimum currency area bites back*. Notenstein Academy White Paper Series. https://www.notenstein.ch/sites/default/files/publications/white_paper_eichengreen.pdf.
European Commission. 1990. One market, one money: an evaluation of the potential benefits and costs of forming an economic and monetary union. *European Economy Economic Papers*, 44.
European Commission. 2011. Green paper on the feasibility of introducing stability bonds. COM(2011) 818 final, Brussels, 23 November.
European Commission. 2012. A blueprint for a deep and genuine economic and monetary union launching a European debate. COM(2012) 777 final, Brussels, 28 November.
European Commission. 2013a. Towards a deep and genuine economic and monetary union: ex-ante coordination of plans for major economic policy reforms. COM (2013) 166 final, Brussels, 20 March.
European Commission. 2013b. Towards a deep and genuine economic and monetary union: the introduction of a Convergence and Competitiveness Instrument. COM (2013) 165 final, Brussels, 20 March.
European Council. 2012a. Towards a genuine economic and monetary union: report by President of the European Council, Herman van Rompuy. EUCO120/12, Brussels, 26 June.
European Council. 2012b. Towards a genuine economic and monetary union. Brussels, 5 December.
Favero, C., and A. Missale. 2012. Sovereign spreads in the euro area: Which prospects for a eurobond? *Economic Policy* 27(4): 231–273.
Fuest, C., and A. Peichl. 2012. *European fiscal union: what is it? Does it work? And are there really 'No Alternatives'*? IZA Policy Paper, 39. Bonn: IZA.
German Council of Economic Experts. 2012. *After the euro area summit: Time to implement long-term solutions*. Special Report 2012. http://www.sachverstaendigenrat-wirtschaft.de/fileadmin/dateiablage/download/publikationen/special_report_2012.pdf.
Gros, D., and T. Mayer. 2010. *How to deal with sovereign default in Europe: Towards a Euro(pean) monetary fund*. CEPS Policy Brief no. 202. Brussels: CEPS.
Honkapohja, S. 2014. The euro crisis: A view from the North. *Journal of Macroeconomics* 39: 260–271.
House of Lords. 2014. *Genuine economic and monetary union' and the implications for the UK*. European Union Committee 8th Report of Session 2013–14. London: The Stationery Office.
Howarth, D., and L. Quaglia. 2013. Banking union as holy grail: Rebuilding the single market in financial services, stabilizing Europe's banks and 'completing' economic and monetary union. *Journal of Common Market Studies*, Annual Review 2012: 103–123.
Krugman, P. 2012. Revenge of the optimum currency area. In *NBER macroeconomics annual,* ed. D. Acemoglu, J. Parker, and M. Woodford, vol. 27.
Lane, P.R. 2012. The European sovereign debt crisis. *Journal of Economic Perspectives* 26(3): 49–68.
Leblond, P. 2014, forthcoming. The logic of a banking union for Europe. *Journal of Banking Regulation* 16.
MacDougall, G.D.A. 1977. *Report of the study group on the role of public finance in European integration*. Luxembourg: OOPEC.
McKinnon, R. 2004. Optimum currency areas and key currencies: Mundell I versus Mundell II. *Journal of Common Market Studies* 42(4): 689–715.
Mongelli, F. 2013. *The mutating euro area crisis: Is the balance between 'sceptics' and 'advocates' shifting*? ECB Occasional Paper No. 144.
Muellbauer, J. 2013. Conditional eurobonds and the euro sovereign debt crisis. *Oxford Review of Economic Policy* 29(3): 610–645.

Obstfeld, M. 2013. *Finance at center stage: Some lessons of the euro crisis*. European Economy Economic Papers No. 493.
Pâris, P. and Wyplosz, C. 2014. *PADRE: Politically acceptable debt restructuring in the Eurozone*. Geneva Special Report on the World Economy, vol. 15. London: CEPR.
Pisani-Ferry, J. 2012a. The known unknowns and the unknown unknowns of European monetary union. *Journal of International Money and Finance* 34: 6–14.
Pisani-Ferry, J. 2012b. *The euro crisis and the new impossible trinity.* Bruegel policy contribution, Issue 2012/01. Brussels: Bruegel.
Pisani-Ferry, J., and G. Wolff. 2012. *The fiscal implications of a banking union*. Bruegel policy brief. Brussels: Bruegel.
Sawyer, M. 2013. Alternative economic policies for the economic and monetary union. *Contributions to Political Economy* 32(1): 11–27.
Schulmeister, S. 2013. *The European Monetary Fund: A systemic problem needs a systemic solution*. Revue de l'OFCE, 2013/01 (no. 127), 389–424.
Tommaso Padoa-Schioppa Group [TPSG]. 2012. *A roadmap towards fiscal union in Europe*. Paris: Notre Europe.
Vallée, S. 2014. From mutual insurance to fiscal federalism: Rebuilding the economic and monetary union after the demise of its Maastricht architecture. *International Economics* 138: 49–62.

George, Henry (1839–1897)

Mason Gaffney

Keywords

Advances; Barriers to exchange; City planning; Clark, J. B.; Common property; Excise taxes; Fabianism; George, H.; Hotelling, H.; Irish land question; Land markets; Land tax; Libertarianism; Lorenz curve; Marginal cost pricing; Market failure; Mixed economy; Public utility pricing; Scarcity; Specialization; Speculation; Supply side economics; Taxation of income; Urban economics; Urban planning; Urban transportation economics; Veblen, T.; Vickrey, W. S.; Wages fund; Wages tax; Wicksteed, P. H.

JEL Classifications

B31

Henry George was by turns sailor, prospector, printer, reporter, San Francisco newspaper editor and publisher, orator and political activist before closeting himself to write on political economy. His *Progress and Poverty* (1879) electrified reformers, catapulted him to fame and began a worldwide movement for land reform and taxation, opening to George an extraordinary career in radical politics. Returning from Ireland as reporter for *The Irish World of* New York he was lionized by Irish-New Yorkers for his stand on the Irish land question. With ethnic, union and socialist backing he formed the United Labour Party and ran for mayor of New York in 1886, nearly winning.

He toured Britain and won over the Radical-Liberals, and then toured Australia as a folk hero. At home he was courted by Democrat and later by Populist leaders. He died in 1897 while running again for New York mayor, but his followers rose in and helped shape the Progressive movement which dominated the next 20 years. His name has become a byword for ideas and policies he espoused.

George is best known today for *Progress and Poverty* (1879). Eloquent, timely and challenging, it soon became and remains the all-time best-seller on economic theory and policy.

George defines 'The Problem' as increase of want with increase of wealth. Dismissing Malthusian fatalism as merely a device to rationalize privilege, George attributes low wages and unemployment rather to artificial scarcity of land and barriers to free exchange. Artificial scarcity results from unequal dispensation of public lands, concentration and 'speculation'. George's speculation is pervasive market failure endemic to land, which failure he attributes to holding for the unearned increment.

George proposed to raise the *ad valorem* property tax rate on bare land (broadly defined as all natural opportunities), thus socializing rent without excess burden. He would remove other taxes, calling them barriers to commerce, employment

and capital formation. The cash drain of the *ad valorem* tax, while neutral at the margin, would move and lubricate the land market as a whole, forcing land into full use. Observation persuaded him that otherwise speculation overrode the incentives to use land fully.

Release of hoarded lands would open wider opportunities for both labour and man-made capital. His overriding concern was for labour, but he saw capital mainly as a form of labour, produced by labour, complementing labour. So in an era before payroll taxes it was actually capital and commerce he sought to untax for the benefit of labour – a preview of 'business Keynesianism' in W. Heller's production of Camelot.

George did not see investment employing labour, but labour producing capital, a difference that to him was more than a nuance. While admiring Quesnay he never absorbed the Physiocratic idea of '*avances*'. Instead he attacked its English derivative, the wages-fund theory with its advances of subsistence, a concept he rejected as condescending to labour. He developed no concept of economic circulation, of either capital or spending. He lacked a good capital theory, belittling Austrian interest theory and botching his own. These faults narrowed the effective scope of his otherwise seminal work and ultimately limited its influence, which is still wide and sustained but mainly outside the macroeconomic field it addressed.

His programme would level barriers to exchange and specialization and production and synergy. These include spatial barriers forced by land speculation (for example, scattered settlement and urban sprawl); fiscal barriers like excise and wage taxes; and social barriers from unequal wealth and contempt for workmanship, which he (like Veblen) traced to the influence of privilege and unearned wealth. This 'true free trade' would unleash technological, scientific, cultural and spiritual development in a more egalitarian and moral society organized around a perfected market mechanism.

George drew on earlier thinkers: Quesnay, Smith, Ricardo, Spencer, and Mill. And he contributed much to later thinking.

George was system-minded and sought to unify the laws of production and distribution in a coordinated harmonious system. His theoretical framework is an early adumbration of the marginal productivity theory of wages, which he integrates with Ricardo's rent law. J.B. Clark was a nemesis, and P. Wicksteed a friend, but both were formalizing insights from George.

Although best known as a deductive thinker, the journalist was also an observer with statistical intuition. In debate with Francis Walker on 'The March of Concentration' in farming, George anticipated Lorenz's method of analysing size distributions and goaded the US Census into publishing farm size data in that form.

George wanted radical redistribution but without revolution. He pioneered the idea that taxation, properly crafted, can redistribute wealth without damage to the market. His influence on Fabianism was early and wide; also on American reformers like Tom L. Johnson, Upton Sinclair, John R. Commons and Norman Thomas. The modern 'mixed economy' is in the Georgist spirit of reform within traditional forms.

Continued heavy reliance on real estate taxation in Canada and the United States, with separate assessment of land value, reflects George's influence, as do the inclusion of land rents and gains in the income tax base, and the efforts of Lloyd George, Asquith and Snowden to introduce national land taxes in Britain.

Free provision of public goods, social dividends, and marginal-cost pricing for urban mass transit and utilities are vintage Henry George. H. Hotelling and W. Vickrey have acknowledged their debt.

The optimistic 'economics of abundance' idea owes much to George. The prevailing 'dismal' economics was a science of choice where all the choices were bad and leaders could only call for more sacrifices. George promised full employment at higher wages by unlocking natural opportunities now held in speculation. Needed capital would be formed in the very process of making jobs, an idea pervading Keynes. Social synergy would produce a surplus that spills over into higher land rents, a 'free lunch' that government may tap in lieu of taxes that penalize and abort useful activity.

George lives too in urban economics and city planning. George's emphasis on the synergistic gains from urban linkages, and the wastes of sprawl caused by failure of the land market anticipates much of planning doctrine. Ebenezer Howard is an obvious link: his 'Garden City' presupposed Georgist taxation to move the land market.

The idea that environment is a common heritage for future generations in pure Georgism. 'Spaceship Earth', common property, and rights of the unborn are his very phrases.

As to economic development, the economists are legion who have recommended a 'dose of Henry George' to help LDCs take off, and some, like Taiwan, belatedly following the counsel of the Georgist Dr. Sun Yat-sen, have taken the dose with good results.

On the conservative side, George was a pioneer of tax limitation, insisting that land rent set an upper limit on government spending. The resurgence of libertarianism and supply-side economics may set a new stage for George, whose programme was mainly oriented to increasing production in the private sector. Religion in politics should not threaten George, who unabashedly presented economic policy as an implementation of religious ideals.

George's blend of radicalism and conservatism can puzzle one until it is seen as a reconciliation of the two. The system is internally consistent but defies conventional stereotypes.

See Also

► Land Tax

Selected Works

Complete works. Garden City: The Fels Fund; also other publishers.

Bibliography

Andelson, R., ed. *Critics of Henry George*. London: Associated University Presses.
Barker, C.A. 1955. *Henry George*. New York: Oxford University Press.
Cord, S.B. 1965. *Henry George: Dreamer or realist?* Philadelphia: University of Pennsylvania Press.
de Mille, A.G. 1950. *Henry George: Citizen of the World*. Chapel Hill: University of North Carolina Press.
Geiger, G.R. 1933. *The philosophy of Henry George*. New York: Macmillan.
George, H., Jr. 1900. The life of Henry George. Reprinted, New York: Robert Schalkenbach Foundation, 1943.
Lawrence, E.P. 1957. *Henry George in the British Isles*. East Lansing: University of Michigan Press and Vanguard Press.
Post, L.F. 1930. *The prophet of New York*. New York: Vanguard Press.
Sawyer, R.A. 1926. *Henry George and the single tax: A catalogue of the collections in the New York public library*. New York: New York Public Library.
The Standard (weekly). 1887–92. New York.

Georgescu-Roegen, Nicholas (1906–1994)

Stefano Zamagni

Keywords

Georgescu-Roegen, N.; Integrability; Transitivity; Non-substitution theorem; Input–output analysis; Separating hyperplane theorem; Regime switching; Bioeconomics; Entropy law

JEL Classifications

B31

Born in Constanza, Rumania, on 4 February 1906, Georgescu-Roegen obtained his first degree in mathematics in 1926 from the University of Bucharest. He then went to Paris where, under the supervision of E. Borel and G. Darmois, he received in 1930 the doctorate in mathematical statistics. In October of the same year he moved to London to pursue further research with K. Pearson. By 1932 Georgescu was Professor of Statistics at the University of Bucharest. His life was inextricably bound up with the social and political events of his country, which explains the emergence of his interest in economics and his

consequent decision to spend a two-year 'apprenticeship' (1934–6) at Harvard where he was able to work closely with Schumpeter. In 1937 he returned to Rumania, where he combined an active academic career with increasing responsibilities in public institutions. In February 1948 he fled from his country and, after a short stay at Harvard, was appointed professor at Vanderbilt University, where he remained until his retirement in 1976.

Georgescu-Roegen's scientific work is notable for an early phase centred around consumer theory, input–output analysis and production theory at large, and a later phase mainly devoted to growth modelling, methodological issues and the ambitious attempt to develop a 'bioeconomic' approach to economic thinking. The early phase is well represented by his 1936 classic article on consumer theory and his 1954 famous paper on 'Choice, Expectations and Measurability'. In the former article, which deals with the 'mysterious' problem of integrability in the theory of demand, one finds two major results: the demonstration that the integral varieties do not necessarily coincide with the indifference varieties – whence the distinction between mathematical integrability and economic integrability – and the demonstration that the two kinds of varieties come to the same thing in the presence of the postulate of transitivity of preferences. The latter essay, focusing on the non-existence of the indifference map of the consumer as a consequence of the pervasiveness of lexicographic ordering of preferences, allowed him to prove what he called the 'ordinalist fallacy' and to inquire about the origin and implications of probabilistic preferences, a subject that is at the very frontiers of economics even today.

On the other front, three contributions are particularly noteworthy. In Georgescu (1951a) we find the first and the most general statement of the celebrated non-substitution theorem: justifying the separation of scale and composition in linear multisectoral models, the theorem provides a theoretical underpinning and analytic rationale for the consistency of input–output analysis. The (1951b) paper offers the first 'geometric' proof of the existence of a von Neumann's equilibrium by using the separating hyperplane theorem – a theorem that was to enter the toolbox of the economist. In his (1951c) essay, Georgescu challenged the two most intractable problems in macrodynamics – nonlinearities and discontinuities – providing, on the basis of an innovative application of the theory of relaxation oscillations, a fundamental result for investigations of regime switching.

The later phase begins with the 1966 famous methodological essay containing Georgescu-Roegen's critique of standard economics for having reduced the economic process to a mechanical analogue and a proposal of a new alliance between economic activity and the natural environment – what later would become his 'bioeconomic programme'. The key to such a project is found in the entropy law ('the most economical of physical laws'), which brought Georgescu to inquiry on the fundamental relation between mankind's existence and its environmental dowry. This problem prompted him to step over the fence of economics into thermodynamics, where he formulated a new law (the 'fourth law'): the impossibility of the perpetual motion of the third kind defined as a *closed* system that could perform work at a constant rate indefinitely. The implications for economics of this line of thinking and in particular of his strong rebuttal of the 'energetic dogma' ('only energy matters') are nicely developed in his 1971 and 1976 books. In this last book, Georgescu lays the foundations of a new approach to production theory: the 'flow-fund' model as a radical alternative to both the production function model and the activity analysis model, models whose main drawback lies in their inability to tackle properly the time element in the productive process.

The long introductory essay (145 pages) that Georgescu wrote in 1983 for the English edition of Gossen's *The Law of Human Relations* is not simply a splendidly written intellectual biography, showing the depth and breadth of his economic culture, but it contains also a restatement in modern analytical terms and an expansion of Gossen's theory of economic behaviour.

Georgescu-Roegen was one of those rare scientists able to couple a remarkable expertise in their specific field with a philosophical bent of mind. In this sense he was a true Renaissance man, which perhaps helps to explain the generalized *fin de non recevoir* of the profession with respect to his critical message, the message of a scholar who cannot be identified with any single school of economic thought and whose intellectual endeavour is best seen as a major contribution to the shifting of the frontiers of economic theory and methodology.

Selected Works

1936. The pure theory of consumers' behavior. *Quarterly Journal of Economics* 50: 545–593.
1951a. Some properties of a generalized Leontief model. In *Activity analysis of production and allocation,* ed. T.C. Koopmans et al. New York: Wiley.
1951b. Relaxation phenomena in linear dynamic models. In *Activity analysis of production and allocation,* ed. T.C. Koopmans et al. New York: Wiley.
1951c. The aggregate line as production function and its applications to Von Neumann's economic model. In *Activity analysis of production and allocation,* ed. T.C. Koopmans et al. New York: Wiley.
1954. Choice, expectations and measurability. *Quarterly Journal of Economics* 68: 503–534.
1966. Some orientation issues in economics. In *Analytical economic issues and problems*, ed. N. Georgescu-Roegen. Cambridge, MA: Harvard University Press.
1971. *The entropy law and the economic process.* Cambridge, MA: Harvard University Press.
1976. *Energy and economic myths: Institutional and analytical economic essays*. Oxford: Pergamon Press.
1983. Hermann Heinrich Gossen: His life and work in historical perspective. In *The laws of human relation and the rules of human action derived therefrom,* trans. from the original German edn of 1854 by R.C. Blitz, ed. H.-H. Gossen. Cambridge, MA: MIT Press.

Gérard-Varet, Louis-André (1944–2001)

Rodolphe Dos Santos Ferreira

Keywords

Adverse selection; Art, economics of; Competitive toughness index; Endogenous fluctuations; Enforcement mechanisms; Expected externality mechanism; Farkas lemma; General equilibrium; Gérard-Varet, L.-A.; Imperfect competition; Industrial organization; Involuntary unemployment; Lagrange multipliers; Market share; Market size; Mechanism design; Moral hazard; Oligopolistic competition; Overlapping generations; P-equilibrium; Pricing schemes; Truth revelation

JEL Classifications

B31

Born on 30 June 1944 at Auxerre, Gérard-Varet studied economics and sociology at the University of Dijon, and prepared his doctorate partly at CORE (Louvain). He taught, as full professor, in Strasbourg, Toulouse and, for most of his academic career, Marseille, where he died on 31 January 2001. He played a quite important role, not only through his teaching and his scientific production, but also as an organizer, in particular as long-term director of his research centre (GREQAM, Marseille) and president of national and international economic associations.

The first set of the theoretical contributions of Gérard-Varet concerns mechanism design, a field in which he began to collaborate in 1973 with Claude d'Aspremont, in the context of a project on cross-border pollution. Starting from the Vickrey–Clarke–Groves mechanism, which ensures that truth revelation by each agent is a dominant-strategy equilibrium, but not that the budget is balanced, they introduced the *expected externality* (or *AGV) mechanism,* which is both truthfully implementable as a Bayesian equilibrium

and budget-balanced. The mechanism was explicit, while requiring independence of agents' beliefs. This condition was considerably generalized by switching from a constructive to an existence proof, based on the Farkas lemma (1979, 1990a). Restrictions on beliefs under adverse selection were also shown to be transposable to stochastic outcome functions in team moral hazard, and to be applicable to two kinds of enforcement mechanisms: enforcement through transfer schemes and enforcement through repetition (1998). More generally, the work of the mid-1970s opened the way to a lifelong research programme.

The second set of the theoretical contributions of Gérard-Varet concerns oligopolistic competition (in partial, general and macroeconomic equilibrium). In 1980, he engaged in the analysis of the macroeconomic effects of significant output market power. He first showed the possibility of so-called involuntary unemployment (in Keynes's sense of persistent unemployment at an arbitrarily low money wage). This results either from the non-existence of a full employment equilibrium (because marginal revenue eventually becomes negative as wages decrease) or, when equilibria are multiple, from a failure to coordinate on that equilibrium (1990b). The early static results were extended to overlapping generation economies and linked to the emerging literature on markup variability as a source of endogenous fluctuations (1995a). The need for a unified treatment of different standard varieties of imperfect competition motivated the formulation of the *P-equilibrium* concept, first applied to an industry or a group of industries (1991), then to the whole economy, in a general equilibrium approach (1997). Strategic agents simultaneously choose price signals in order to manipulate market prices through some *pricing scheme P*, and quantities, required to satisfy market realization constraints. A distinct but related concept of oligopolistic equilibrium was designed later (2007), where producers of elements of a composite good choose price–quantity pairs under two constraints, on market share and on market size. The associated Lagrange multipliers are used to build an index of *competitive toughness*, parameterizing the set of equilibria and appearing as a foundation to the 'conjectural variations' parameter of the empirical industrial organization studies.

The preceding themes by no means exhaust the list of subjects on which Gérard-Varet has made theoretical and applied contributions, often combining features of public economics and industrial organization, of which the economics of visual arts is a good example (1995b).

See Also

► Mechanism Design

Selected Works

1979. (With C. d'Aspremont.) Incentives and incomplete information. *Journal of Public Economics* 11: 25–45.

1990a. (With C. d'Aspremont and J. Crémer.) Incentives and the existence of Pareto- optimal revelation mechanisms. *Journal of Economic Theory* 51: 233–254.

1990b. (With C. d'Aspremont and R. Dos Santos Ferreira.) On monopolistic competition and involuntary unemployment. *Quarterly Journal of Economics* 105: 895–919.

1991. (With C. d'Aspremont and R. Dos Santos Ferreira.) Pricing schemes and Cournotian equilibria. *American Economic Review* 81: 666–673.

1995a. (With C. d'Aspremont and R. Dos Santos Ferreira.) Market power, coordination failures and endogenous fluctuations. In *The new macroeconomics: Imperfect markets and policy effectiveness*, ed. H. Dixon and N. Rankin. Cambridge: Cambridge University Press.

1995b. On pricing the priceless: Comments on the economics of the visual art market. *European Economic Review* 39: 509–518.

1997 (With C. d'Aspremont and R. Dos Santos Ferreira.) General equilibrium concepts under imperfect competition: A Cournotian approach. *Journal of Economic Theory* 73: 199–230.

1998 (With C. d'Aspremont.) Linear methods to enforce partnerships under uncertainty: An

overview. *Games and Economic Behavior* 25: 311–336.
2007. (With C. d'Aspremont and R. Dos Santos Ferreira.) Competition for market share or for market size: Oligopolistic equilibria with varying competitive toughness. *International Economic Review* 48(3) (forthcoming).

German Economics in the Early 19th Century

K. Tribe

Abstract

During the 18th century 'cameralism' became a recognized part of the introductory curriculum of (mainly northern, Protestant) German universities. The French Revolution, the Napoleonic occupation, and Kant's new Critical Philosophy together swept established doctrine aside during the final decade of the century, allowing a reoccupation of the university curriculum by 'modern economics'. Jean-Baptiste Say had more direct influence on this than Adam Smith's *Wealth of Nations.* When combined with a post-Critical emphasis on human needs, this directed German writings away from the English emphasis on value and distribution and laid the foundation for a new 'marginalist' economics in the 1870s.

Keywords

Cameralism; Engels, F.; German economics in the early 19th century; Hermann, F.; Hufeland, G.; Jakob, L.; Marx, K.; Menger, C.; Musgrave, R.; *Nationalökonomie*; Physiocrats; Rau, K.; Sartorius, G.; Say, J.-B.; Schumpeter, J.; Smith, A.; Steuart, J.; von Soden, J.

JEL Classifications

B1

Nationalökonomie emerged in early 19th century Germany as a new doctrine which took human needs and their satisfaction to be the first principle of economic analysis.

Cameralism had become in mid-18th century Germany a regular part of the university curriculum in German universities, taught in faculties of philosophy to future state officials. By the 1820s this function had been modified, with lectures on economics becoming a compulsory part of the law curriculum, as elsewhere in Continental Europe. This institutional development coincided with the emergence of a new economics which displaced the older *Cameralwissenschaften,* together with their focus on state finance and national wealth. The older teaching was pushed to one side, and its place in the lecture room taken by the new principles and doctrine of *Nationalökonomie.* This joint development was not a direct outcome of the Revolutionary and Napoleonic Wars (although these certainly caused a great deal of physical disruption to universities), nor a consequence of the new republican ideas fostered by these wars and by the French Revolution itself (although these certainly played some role). Nor did it follow directly from the diffusion of Adam Smith's teaching, although Smith did in the early 1800s become a major point of reference in discussion of economic principles. This internal transformation in the teaching of economics in German universities resulted instead from an assault upon the older, eudaemonistic natural law tradition by converts to Critical (Kantian) Philosophy. This process was just gaining momentum around 1795; by 1805 the transformation was all but complete. Although textbooks in the cameralistic tradition still appeared, and, it can be assumed, professors continued as always to read out their old lectures, the teachings of Smith and Say now found a definite place within the university, as part of a new *Fach, Nationalökonomie.*

This new science drew on a range of sources, but remained quite distinct from the political economy being developed at the time in Britain by James Mill, Robert Malthus and David Ricardo. It shared neither their emphasis upon distribution between agents defined by the process of production, nor the peculiarly English

preoccupation with value and its measurement. As we shall see, German writers were, directly or indirectly, influenced by the work of Jean-Baptiste Say rather than by any English political economist. Especially important was Say's tripartite schema of production, distribution and consumption, and his argument that production produced neither 'things' nor value, but *utilities.*

A direct line can therefore be traced from this early *Nationalökonomie* to the work of the early Austrian economists. When in 1871 Carl Menger published his *Grundsätze der Volkswirthschaftslehre,* he defined as 'goods' 'utilities . . .related to the satisfaction of human needs' (1871, p. 2). A long footnote was appended to this statement, beginning with Aristotle's conception of goods, proceeding on through Forbonnais, Le Trosne and Say; and listing as the first relevant German authors Soden, Jakob and Hufeland. These three writers were the principal architects of the new *Nationalökonomie.* Among them, Jakob's definition of a good is the most pithy: 'Everything that serves the satisfaction of human needs' (1805, §. 23). Jakob was also Say's German translator (1807), his own textbook of 1805 clearly following the organization and argument of Say's *Traité d'économie politique.* And, like Schlözer (1805, §. 12), Jakob made a clear distinction between state and economy, or politics and economics.

The Definition of *Nationalökonomie*

'German economics' has always been primarily an academic rather than a public or popular discourse, but during the 18th century it had been not uncommon for teachers of cameralism to be shared with the practical world of state administration. Soden was one of the last representatives of this tradition – he was never a university teacher, but spent his working life in state administration before retiring in 1796 to his estate so that he might devote himself to literary pursuits. Prompted by what he saw as the lack of system in Adam Smith's *Wealth of Nations,* a new translation of which had just appeared (1794–6), Soden set out to provide a more systematic basis for the new science, defining it as:

> . . .the *Natural Law* of *sociable* mankind with respect to the maintenance and promotion of its physical welfare, and in the same way that the Law of Nations outlines the laws according to which *nations,* in the reciprocal condition of co-existence, must adhere in every respect; so *Nazional- Oekonomie* provides the principles which . . . must be adhered to, such that every member of every nation achieves the highest possible degree of physical welfare, and maintains this position. (1805, pp. v, vi)

The leading principle of *Nationalökonomie* was then described as '. . . the highest perfection of the physical condition of sociable mankind' (1805, p. 14), underscoring the impact of the contemporary intellectual fashion for Critical Philosophy beyond the confines of philosophy and ethics.

Gottlieb Hufeland, the third writer named by Menger, was, like Jakob, a professor of law who had become a 'Kantian' – and both, like Soden, found their way to Adam Smith via Kant's new philosophy. Hufeland's *Neue Grundlegung der Staatswirtschaftskunst* began with a review of the relative merits of James Steuart and Adam Smith as systematic theorists of economic life. Smith's *Wealth of Nations* had quickly been translated into German, but found at first little resonance among university professors, although, like the Physiocrats, Smith was more widely read by lay members of local literary and economic societies than by university scholars. The German 'Smith reception' proper began with the publication of the second translation in the mid-1790s, coinciding with the developing wave of enthusiasm for Critical Philosophy. And as Hufeland noted in his introductory remarks, it was only towards the end of the 1790s that there were clear signs that Smith had been read *and* understood (1807, n.p.).

Soden and Jakob, observed Hufeland, had dubbed their new field of study *Nationalökonomie,* and, while he did not find this very objectionable, he suggested that it would be 'better and clearer' to use a German expression, *Volkswirthschaft* – which is indeed the root term that became generally accepted about a century later. This latter expression was more suitable, he thought, because it expressed a clear distinction with respect to *Staatswirthschaft*, the generic term that cameralists had used to describe the domain of economic life in

an intellectual system where no distinction was made between 'state' and 'society'. The problem with the German word *Wirthschaft*, he noted, was that it implied a governing person – invoking the Aristotelian head of household on the one hand, and the more down-to-earth figure of the farmer or inn-keeper on the other *(Landwirt,* or simply *Wirt).* Such a figure was absent from the *Volkswirthschaft* 'where many thousands pursue their economic life' *(wirthschaften)'* (1807, p. 14). This he later clarified as a 'sphere of goods', goods being defined as any medium for the realization of human purposes; hence, the 'sphere of goods' was a domain of autonomous human economic activity independent of state action (1807, pp. 17–18, 116).

Jakob and the Architecture of *Nationalökonomie*

Of the writers introduced so far, Ludwig Heinrich Jakob was the most influential in recasting German economics around the conception of human need. He began teaching a course on 'Political Economy and State Economy according to Sartorius' in 1801. Up to this time he had been preoccupied with the creation of a new natural law based on critical principles, exemplified by his *Philosophische Rechtslehre oder Naturrecht* of 1795. We can better see how this new natural law contributed to the reshaping of cameralism if we consider the structure of the book that Jakob first used as a textbook, Sartorius's *Handbuch der Staatswirthschaft zu Gebrauche bey akademischen Vorlesungen* (1796). It was a standard requirement that professors select a textbook to which their lectures were directed, and quite normal for a new course of lectures to be developed as a commentary on this text. It had also become established practice that each lecturer found the existing texts in some way unsuited to his purposes, and so from this commentary there would develop a new text which became in turn the assigned text. If simple enthusiasm for the writings of Adam Smith had been sufficient to bring about the demise of cameralism and its replacement by *Nationalökonomie,* then we might reasonably expect Sartorius to have played a key role in this, for he was one of the first and most articulate Smithians. But while many late cameralistic writers recognized that Smith's *Wealth of Nations* was an important work, none of them contributed significantly to the new conception of human need, economic activity and welfare that was to survive as the core of German economics for more than a century.

Sartorius's textbook carries the subtitle *Nach Adam Smiths Grundsätzen ausgearbeitet,* for it is principally a condensed version of *Wealth of Nations,* about 40,000 words long, based on lectures that Sartorius had delivered in Göttingen since 1791 as a Privatdozent in the philosophy faculty. This, therefore, gives us an idea of what Sartorius taught, and also what argumentative opportunities this text presented to Jakob in his own initial lectures. Sartorius's presentation is brisk: Smith's 'Introduction and Plan of the Work' is dealt with in 16 lines, and by the fifth page we have already arrived at Book I, Ch. V. Books I and II of *Wealth of Nations* are dispatched in 90 pages of summary, each paragraph corresponding to a chapter or a part of a chapter in the original. Discursive sections of *Wealth of Nations* are reduced to bare propositions; importantly, the argument concerning the human propensity to exchange is suppressed entirely. Once the summary reaches the end of Smith's Book II, Sartorius inserts his own section which summarizes the principles of Books III–V as if they were part of a cameralistic treatise on *Staatswirtschaft:* the title runs 'Of State Economy, or the Rules which the Government of a State must Pursue, so that Individual Citizens might be placed in the Position of being able to Create for Themselves a Sufficient Income, as well as Providing the Same for Public State Expenditures'. Here, although freedom is the means, a eudaemonistic conception of welfare is the objective. This shift towards an older German conception of the state, its tasks and objectives is continued into the treatment of public finances.

For all his admiration of Smith during the 1790s, Sartorius presented a brisk précis of *Wealth of Nations* rather than a summary of, or commentary upon, its leading principles. The

characteristic emphasis that we have seen in Jakob, Hufeland, Schlözer and Soden upon human needs, upon goods as means for the satisfaction of such needs, and upon the economy as the domain within which human individuals sought to maximize their satisfaction of need finds no place here. Two years later Say's *Traité* was published, and this would prove a far more promising avenue through which these concerns, springing from Critical Philosophy, could be brought to bear upon economizing activity.

Jakob notes in the Preface to his *Grundsätze* that he had used Sartorius's *Handbuch* for some years, but that he had come to the view that some of Smith's ideas were obscured by the form of presentation adopted. He proceeds to redefine the state and its affairs in a manner that denies it a decisive role in the formation and distribution of wealth, analogously to the manner in which Say had clearly separated politics from political economy. The expression 'State', Jakob argues, can be used to refer only to *public* affairs; state property is therefore merely a part of national property *(Volksvermögen),* separate from it and for use in pursuit of public and common ends.

> *Staatswirthschaftslehre* can be in fact nothing other than *financial science* or *Policey,* insofar as care for public order is part of good public economy. (1805, p. vi)

How does this intent translate into the principles of *Nationalkönomie,* and what relationship is established to political economy on the one hand and the cameralistic sciences on the other?

The most striking initial feature of Jakob's book is its plan of organization. At the end of the 'Introduction', he states that *Nationalökonomie* deals with three principal issues:

1. The formation and increase of national wealth.
2. The principles of the most advantageous distribution of national wealth among the members of society.
3. The consumption of national property and the various effects of the same. (1805, p. 12 §. 20)

Or, in other words, the trinity of production, distribution and consumption introduced by Say in the second edition of his *Traité* (1814). Say had already stated in his first edition that Smith distinguished politics as the science of legislation from a political economy dealing with the formation, distribution and consumption of wealth (1803, vol. 1, pp. i–ii); but this first edition was divided into five books, dealing in turn with production, money, value, revenue and consumption. Jakob, by contrast, not only stated that *Nationalökonomie* dealt with the production, distribution and consumption of wealth; his book is divided up in this way too. Viewed from this perspective, the sequence of chapters in Jakob's 1805 textbook resembles more closely the order in which material is treated in Say than it is in Smith.

Karl Heinrich Rau and the Systematization of *Nationalökonomie*

Jakob, Soden, Hufeland and Schlözer, more or less simultaneously and independently, created a new conception of economic life separate from the work of state administration, which had hitherto been thought to provide a necessary framework for the orderly conduct of production and consumption. This did, however, remain very general in outline, and in many cases it was simply taught alongside the traditional practical areas of economic administration, such as agriculture and forestry, finance, and botany. In 1826 Karl Heinrich Rau, since 1822 a professor at Heidelberg, published the first volume of a textbook that was to end this equivocal state of affairs. Three volumes of his new *Lehrbuch der politischen Oekonomie* – Rau chose to revert to a name for the subject generally accepted outside Germany and more recognizable to French or English readers – were published between 1822 and 1837. The work ran into many editions, the last revised edition appearing in 1876. Here again there is a clear line of connection to Austrian economics, for Menger drafted the first outline of his *Grundsätze* in the margins of his 1863 edition of Rau (Menger 1963).

The first volume of the *Lehrbuch* deals with 'Die Volkswirthschaftslehre' – 'those characteristic laws which can be perceived in the economic

activities of peoples regardless of the intervention of government' (1826, p. x). It begins by making a clear distinction between private and public economics. 'Private economics' is composed of the rules governing the optimum satisfaction of needs through the acquisition, maintenance, and use of material goods. 'Public economics' by contrast deals with the satisfaction of needs by the allocation of material goods on the part of the state – it has a strictly redistributive character, recirculating goods produced in the 'private economy' using revenues derived from taxation. Whereas the *Volkswirthschaft* is conceptualized by the individual pursuit of self-interest, there are general aims of the state that the individual cannot attain unaided, and so the role of government in the economy is to intervene to ensure that these general aims are secured. Rau's account does not have the kind of internal theoretical structure that readers of English political economy might anticipate. Although he announces his intention to develop a theory of economic forces based upon natural laws, the volume merely enumerates economic objects without regard to their mutual relationship in production, distribution and consumption. The second volume, first published in 1828, is devoted to economic welfare. Although this has affinities with the older cameralistic teaching, Rau here consistently distinguishes between the wealth of an individual and the wealth of a people, only the latter being the proper object of economic analysis. The function of the state is strictly limited to the facilitation of individuals' desire to better their conditions, through educational provision and the promotion of commercial enterprise. Again, however, once past the initial principles the work becomes a review of specific measures fostering enterprise or removing hindrances to individual enterprise. Thus, we read under the heading 'Promotion of Exchange or Encouragement of Trade' about newspapers, fairs, weights and measures, money, roads, railways, canals, and bridges. Later we can read that savings and insurance are to be promoted, and gambling restricted. The proper employment of state expenditure and the relation of taxation to such employments are dealt with in the third volume, devoted to 'financial science'. The first edition appeared in 1832, and Rau was revising the text for a sixth edition at the time of his death in 1870. Shortly beforehand he had suggested to his family that the work be taken over by Adolph Wagner, who published in 1872 his own revised version of Rau's treatment of financial science, which in successive editions became the standard textbook on finance up to and beyond the turn of the century.

Schumpeter's *History of Economic Analysis* (1954, p. 503, n. 2) devotes no more than a dozen lines to Karl Heinrich Rau, dismissing his textbook as adequate for teaching but of little further interest. But, as noted above, it was with this textbook that Menger started to draft his *Grundsätze,* while the link to Wagner reaches on to Richard Musgrave's work on public finance (he completed his *Diplom Volkswirt* in Heidelberg in 1933) and thence to Buchanan's conception of public goods. And Rau is also linked to the final writer considered here, Friedrich Benedict Hermann, who graduated from the University in Erlangen in 1823 only one year after Rau left for the chair in Heidelberg.

Towards Post-classicism

Hermann's *Staatswirthschaftliche Untersuchungen* of 1832 sketched a clear relation between the supply of and demand for economic goods as formative of market prices. His introductory discussion identifies the level of profit and the relation of profit to wages as the most difficult area of economic analysis, given a rigorous treatment by Ricardo but still in need of much refinement. This immediately establishes an approach far more theoretical in character than was at the time usual for German writers on economics But, while it is true that Hermann's work looks forward to the kind of discussion of price formation that we later find elaborated in Mangoldt (1863), it is clear that Hermann shares the preconceptions of all the writers outlined above. His account of basic principles begins with the definition of a good as anything satisfying human need, an 'economic good' being acquired through sacrifice of labour or money; and the book ends with the statement

that consumption is destruction of use value, a conception drawn from Jean-Baptiste Say.

The initial discussion of need and its satisfaction reviews their treatment in James Steuart's 1767 *Principles* and Say's 1828 *Cours,* arguing that use value is the main feature of a good because of its capacity to satisfy needs. This does not, however, prevent Hermann from developing an analysis of price formation in which the price level for a particular good is made dependent upon the relation of demand and supply or, what is much the same thing, the relation between the number of sellers and the number of buyers, which echoes Jakob's account of prices and effective demand, with the addition of the term 'equilibrium' to describe the point where '. . .goods are demanded and supplied in the same quantities' (1832, p. 67). Given a basic cost which includes the usual rate of interest and entrepreneurial profit, he suggests that if the price falls below cost then capital and talent will move elsewhere; conversely, where the price prevails above cost then new entrepreneurs will be attracted, in turn leading to a steady reduction in the price until once more prices and costs are equalized (1832, pp. 4–5, 67–81). But even elementary arguments with this degree of clarity are rare in the literature of the period.

Conclusion

This account of German economics in the early 19th century has emphasized the way in which economic discourse had long been a part of university teaching. Left out of the above account are those lacking a university background and whose work thus falls outside this tradition, but who have since become noted as important parts of the early 19th century German context. First among these is Adam Müller, whose Dresden lectures of 1808–9 countered the new idea of society as a self-organizing system with a romantic, organic conception of man and the state (1922). This found no general resonance in contemporary economic writing, and was rediscovered later by early 20th century cultural critics of capitalism. The principal contemporary influence attributed to Müller connects him to Friedrich List, but there is little evidence for this supposition. List for his part did briefly teach at Tübingen during 1818 and 1819, but in 'administrative practice', not 'economics' in even the widest contemporary sense. The arguments that he developed during the 1830s and early 1840s, and for which he has been remembered, developed economic ideas he had discovered in American writing during the 1820s, and have no direct relation to German economics in this period. The reputation of Karl Marx, who as a law student during the late 1830s in Berlin would presumably have been exposed to some lectures in economics, likewise owes nothing to contemporary German economic writings, since his own interest in political economy was first stimulated by Friedrich Engels's enthusiasm for (English) Owenite ideas, was later developed as a critique of the classical economics of Mill and Ricardo, and betrays little knowledge of contemporary German writing in politics and economics.

See Also

- ▶ Cameralism
- ▶ Musgrave, Richard Abel (1910–2007)

Bibliography

Hermann, F. 1832. *Staatswirthschaftliche Untersuchungen*. Munich: Anton Weber.

Hufeland, G. 1807. *Neue Grundlegung der Staatswirthschaftskunst*, vol. 1. Gießen: Tasche und Müller.

Jakob, L. 1805. *Grundsätze der National-Oekonomie oder National- Wirthschaftslehre*. Halle: Ruffsche Verlagshandlung.

Menger, C. 1871. *Grundsätze der Volkswirthschaftslehre.* Gesammelte Werke. Bd.1. Tübingen: J.C.B. Mohr (Paul Siebeck), 1968.

Menger, C. 1963. *Carl Mengers erster Entwurf zu seinem Hauptwerk 'Grundsätze'*. Tokyo: Library of Hitotsubashi University.

Müller, A. 1922. *Die Elemente der Staatskunst,* 2 vols. Jena: Gustav Fischer.

Rau, K.H. 1826–37. *Lehrbuch der politischen Oekonomie,* 3 vols. Heidelberg: C. F. Winter.

Rau, K.H. 1872. *Lehrbuch der politischen Oekonomie,* ed. A. Wagner, Bd. III 1, 2, 6th edn. Leipzig: C. F. Winter.

Sartorius, G. 1796. *Handbuch der Staatswirthschaft zu Gebrauche bey akademischen Vorlesungen, nach*

Adam Smith's Grundsätzen ausgearbeitet. Berlin: J. F. Unger.
Say, J.-B. 1803. *Traité d'économiepolitique,* 2 vols. Paris: Deterville
Say, J.-B. 1807. Abhandlung über die Nationalökonomie, 2 vols. Trans. L. Jakob. Halle: Ruffsche Buchhandlung.
Say, J.-B. 1814. *Traité d'économie politique,* 2 vols, 2nd edn. Paris: A.-A. Renouard.
Schumpeter, J. 1954. *A history of economic analysis*. London: George Allen and Unwin.
Smith, A. 1794–6. *Untersuchung über die Natur und Ursachen des Nationalreichthums,* 4 vols. Trans. C. Garve, and A. Dörrien. Breslau: Wilhelm Gottlieb Korn.
Streissler, E. 1994. German predecessors of the Austrian school. In *The Elgar companion to Austrian economics*, ed. P. Boettke. Aldershot: Edward Elgar.
Tribe, K. 1988. *Governing economy: The reformation of German economic discourse 1750–1840*. Cambridge: Cambridge University Press, Chapters 8 and 9.
Tribe, K. 1998. Natural law and the origins of *Nationalökonomie:* L. H. von Jakob. In *The rise of the social sciences and the formation of modernity: Conceptual change in context, 1750–1850*, J. Heilbron, L. Magnusson, and B. Wittrock. Dordrecht: Kluwer.
Tribe, K. 2002. The German reception of Adam Smith. In *A critical bibliography of Adam Smith*, ed. K. Tribe. London: Pickering and Chatto.
Tribe, K. 2003. Continental political economy from the Physiocrats to the marginal revolution. In *The Cambridge history of science*, vol. 7, ed. T. Porter and D. Ross. Cambridge: Cambridge University Press.
von Mangoldt, H. 1863. *Grundriß der Volkswirthschaftslehre*. Stuttgart: J. Engelhorn.
von Schlözer, C. 1805. *Anfangsgründe der Staatswirthschaft oder die Lehre von dem Nationalreichthume*, vol. 1. Riga: C. J. G. Hartmann.
von Soden, J. 1805. *Die Nazional-Oekonomie*, vol. 1. Leipzig: Johann Ambrosius Barth.

German Historical School

F. Schinzinger

The German historical school is very closely connected to Romanticism and the rise of nationalism in Germany; it is considered a reaction to English enlightenment and classical economics. This reaction to English classical economics manifested itself in two different ways; by developing different methods and by seeking alternative aims in economic research.

The classical school's deductive method is criticized as being too abstract. The German historical school puts the emphasis on the inductive method. Historians point out that economic development is unique, so there can be no 'natural laws' in economics. The economist can only try to show patterns of development common to different economies. Instead of searching for generally applicable laws, the historical school therefore tried to describe the particulars of each era, society and economy. A rational approach to human behaviour is criticized as being unable to show correctly the amplitude of human motives – these being influenced by non-economic principles, even where economics are concerned.

The aims of economic research were put differently: research for research's sake must be abandoned, it must be seen as a means of achieving sensible economic policy, useful for society. This leads to another aspect of the German historical school: ethics. One of the reasons for the rise of the historical school was the social question, namely the problems arising in Germany in the middle of the 19th century. These led to the belief that free trade was unable to solve problems of industrialization in a country totally different from England. From the ethical point of view the German historians demanded that the state had an important role to play in economic affairs. The historical school can be considered as the beginning of the end of liberal economic policy in Germany.

Friedrich List, considered as a forerunner of the historical school, criticized 'free trade' and put forward the idea that it was the duty of the state to protect the still young German industry from the competition presented by a much further developed English industry. He also suggested that the state should protect the socially weak sections of the population. These ideas arose from the phenomenon of 'Pauperismus' in Germany in the 1830s and 1840s – the poverty of millions of people who were no longer able to find work in agriculture nor in slowly developing industry.

The ever-widening rift between economic theories and experienced reality set off a new direction in economic research. With industrialization

progressing in countries, whose social conditions and economic basis were totally different from those in 18th-century England, it seemed necessary to adapt economic research to changing reality.

An attempt to bridge this rift was made in two ways: on the one hand there was the attempt to find a totally new theory which would be more comprehensive than classical theory; on the other hand there was a tendency to dismiss theory and try to see the depiction of reality, in a historical perspective, as the only sensible aim of economic research. For these reasons the historical school is characterized by the development of statistics and economic history.

It is difficult to discover the general opinions of all the economists of the historical school. Few of their ideas were formulated in a clear, non-ambiguous way. The general ideas common to all of them must be filtered out from their works, and this leads to a subjective interpretation. Generally it can be said that all the economists of the German historical school put forward criticisms of the methods of classical economy, especially of deductive methods – even if some of them used such methods in their own works.

Another point common to them all is their criticism of the classical belief in harmony that results from the individual's knowledge and rational following of his economic advantages. German historians emphasized the non-rational influences which lead to human actions, and they also stressed the fact that the individual is part of a socially unique context, which differs in time and space (e.g. differences between the industrialization of England and Germany in the 19th century).

The German historical school has been divided into two epochs, the older and the younger. The older historical school can be attributed to the 1840s–1870s.

The beginning of the historical school is dated 1843 because the first representative of the school, Wilhelm Roscher, then published his book *Grundriss zu Vorlesungen über die Staatswirtschaft nach geschichtlicher Methode*. In view of this he is seen as the founder of the Historical School. He tried to illustrate classical theory with historical examples and his goal was to use the classical theory as a basis for practical economic policy. He confronted the universal claim of the classical theory with the individuality of each single national economy. Economics as a science should try to find out the interactions between ethical, political and economic phenomena. The most important result of Roscher's work was to put forward the non-economic factors which influence economic life. He tried to find laws of development in economies using the method of comparative induction and comparing different times, peoples, countries and cultures.

The second representative of the German historical school is Bruno Hildebrand. He had a more ambitious programme of research than Roscher. His main, uncompleted work is *Die Nationalökonomie der Gegenwart und Zukunft* (1848). He stresses much more sharply than Roscher the differences between the German historical school and classical economics. For Hildebrand, history is a means of renewing economic research and thought. He tried to show the differences between the economies of different times, people and states. He especially tried to find out the laws of economic development (*Lehre der Entwicklungsgesetze der Völker*) with the help of statistical data. In order to help this research he founded the journal *Jahrbücher für Nationalökonomie und Statistik,* which still exists.

The new method of the historical school is theoretically best illustrated by Karl Knies. His book *Die politische Ökonomie vom Standpunkt der geschichtlichen Methode* (1853) is, from a theoretical point of view, more refined than the books of Hildebrand and Roscher. He also accentuates the need to find a new method in economic research. This new method is somewhat different from what Hildebrand and Roscher advocated. Knies was sceptical about the laws of economic development which Hildebrand tried to discover. For Knies, there are only analogies and not 'laws' of economic development in different peoples; economic thought develops alongside economic conditions.

Deciding which economists to attribute to the younger historical school is a point of controversy, since every German economist at the end of the 19th century was formed by this school.

The head of the younger school was surely Gustav Schmoller, who dominated German economics from the 1870s to the end of the 19th century.

Characteristic of Schmoller and his school is the fact that they do not specifically deny that 'laws' and regularities exist in economic and social life – in some ways they are themselves deterministic when they try to find out these regularities. They wrote a large number of monographs, which can be considered works of economic history. As well as this they found another area of research, the solving of practical problems of the day, especially in the social field.

In economic policies the work of the younger historical school can be characterized by its desire to eliminate the negative results of economic liberalism (especially after the 'Gründerkrise' of 1873), by demanding that the state intervene. Schmoller states that the classical theory is unable to solve the problems of the working classes. The discussion now arises around the question of *how* the state should intervene.

In the field of economic policy the younger historical school had its greatest practical success. The historians were called 'Kathedersozialisten' because most of them were professors. They asked for social laws, insurance against illness, accident, old age and unemployment and founded the 'Verein für Socialpolitik', a forum where these demands were put forward and discussed. The practical result of these demands were the social laws of the 1880s which gave German workers insurance against illness, accident and old age – then unique in Europe.

The younger historical school has found fame through a discussion of methods between Gustav Schmoller and Carl Menger. Menger published in 1883 *Untersuchungen über die Methoden der Sozialwissenschaften und der Politischen Ökonomie insbesondere,* to which Schmoller answered with his article *Zur Methodologie der Staats- and Sozialwissenschaften*.

The books of Menger and Schmoller gave rise to a very polemical discussion about the methods of economic research. Menger defended the deductive method against the historical research work of the historical school. In this fight over method all those aspects which had been brought forward in the discussion of the older historical school arose again – although in a more refined way.

It is difficult to say which of the writers at the end of the 19th century can be counted among the economists of the younger historical school: it has been said that Albert Schäffle belonged to it. Schäffle believed in the compatibility of planned production with individual liberty to consume. These ideas were opposed by Lujo Brentano, also attributed to the younger historical school, who pointed out that it was impossible to have individual consumer freedom while there was a central production plan, because consumer demand was mostly irrational. Adolph Wagner has also been counted among the representatives of the younger historical school. His main works dealt with public finance and he gave the state an important role in directing the course of economy Karl B*ü*cher put forward the idea of stages of economic evolution, which had been discussed since the first half of the 19th century. Werner Sombart, whose major work *Der moderne Kapitalismus* describes the history of capitalism, was influenced by the younger historical school, but cannot be attributed to it, because he later put the accent on very different problems.

The German historical school cannot be understood without knowledge of the economic history of Germany in the 19th century. It is mostly the result of social problems arising from population growth at this time and those emerging with industrialization in Germany. It is also the result of increasing nationalistic feelings in a country divided into more than 39 sovereign states. For the younger historical school the economic crisis of the 1870s was an important departure point in demanding state intervention in economics.

The historical background leads to the fact that apart from many different ways in reacting to classical economics the economists of the historical school had many things in common, which justify their incorporation under the same heading. The main idea is that each economic phenomenon is a product of its social context, having grown historically as the result of a long process.

The historical school was typical for Germany in the 19th century, having little influence

elsewhere. Its view of human behaviour asked for research in the field of social psychology. In France this led to the development of sociology and social history. The younger historical school had some influence in the United States, where institutionalism can be seen as an epoch of American economic thought.

See Also

► Schmoller, Gustav von (1838–1917)

Bibliography

Brentano, L. 1888. *Die klassische Nationalökonomie*. Leipzig.

Brinkmann, C. 1937. *Gustav Schmoller und die Volkswirtschaftslehre*. Stuttgart: Kohlhammer.

Bücher, K. 1893, 1918. *Die Entstehung der Volkswirtschaft. 6 Vorträge.* Pts I and II. Tübingen: Laupp.

Cunningham, W. 1894–5. Why had Roscher so little influence in England? *Annals of the American Academy* 5.

Diehl, K. 1941. *Die sozialrechtliche Richtung in der Nationalökonomie*. Jena.

Eisermann, G. 1956. *Die Grundlagen des Historismus in der deutschen Nationalökonomie*. Stuttgart.

Hildebrand, B. 1848. *Die Nationalökonomie der Gegenwart und Zukunft*, vol. 1. Frankfurt am Main.

Keynes, J.N. 1891. *The scope and method of political economy.* London: Macmillan.

Knies, K.G.A. 1850. *Die Statistik als selbständige Wissenschaft*. Kassel: Verlag der J. Luckhardt'schen Buchhandlung.

Knies, K.G.A. 1853. *Die politische Ökonomie vom Standpunkte der geschichtlichen Methode*. Braunschweig: Schwetschke.

List, F. 1925. *Das nationale System der politischen Ökonomie*, 8th ed. Stuttgart/Berlin.

Marshall, A. 1897. The older generation of economists and the new. *Quarterly Journal of Economics* 11.

Menger, C. 1883. *Untersuchungen über die Methode der Socialwissenschaften, und der politischen Ökonomie insbesondere*. Leipzig: Duncker & Humblot.

Montaner, A. 1948. *Der Institutionalismus als Epoche amerikanischer Geistesgeschichte*. Tübingen: Mohr.

Roscher, W.G.F. 1843. *Grundriß zu Vorlesungen über die Staatswirtschaft nach geschichtlicher Methode*. Göttingen.

Roscher, W.G.F. 1854. *System der Volkswirtschaft. Ein Hand- und Lesebuch für Geschäftsmänner und Studierende*, Die Grundlagen der Nationalökonomie, vol. 1. Stuttgart: Cotta.

Schäffle, A.E.F. 1873. *Das gesellschaftliche System der menschlichen Wirtschaft*, 2nd ed. Tübingen.

Schäffle, A.E.F. 1878. *Enzyklopädie der Staatslehre.* Tübingen.

Schmoller, G. 1888. Die Schriften von Menger und W. Dilthey zur Methodologie der Staats- und Sozialwissenschaften (1883). In *Zur Litteraturgeschichte der Staats- und Sozialwissenschaftén*, ed. G. Schmoller, 275–304. Leipzig: Duncker & Humblot.

Schmoller, G. 1890. *Zur Sozial- und Gewerbepolitik der Gegenwart*. Leipzig.

Schmoller, G. 1897. *Wechselnde Theorien und feststehende Wahrheiten im Gebiete der Staats- und Sozialwissenschaften und die heutige deutsche Volkswirtschaftslehre*. Berlin.

Sombart, W. 1902. *Der moderne Kapitalismus*, 2 vols. Leipzig: Duncker & Humblot.

Sombart, W. 1903. *Die deutsche Volkswirtschaft im 19. Jahrhundert*. Berlin: Georg Bondi.

Sombart, W. 1925. *Die Ordnung des Wirtschaftslebens*. Berlin.

Sombart, W. 1930. *Die drei Nationalökonomien. Geschichte und System der Lehre von der Wirtschaft*. Munich.

Spiethoff, A. (ed.). 1938. *Gustav von Schmoller und die deutsche geschichtliche Volkswirtschaftslehre, Festgabe zur 100. Wiederkehr seines Geburtstages*. Berlin.

Veblen, T. 1919. *The place of science in modern civilization*. New York: Huebsch.

von Böhm-Bawerk, E. 1924. *Historische und theoretische Nationalökonomie*. Jena: Fischer.

Wagner, A. 1895. *Sozialismus, Sozialdemokratie, Katheder– und Staatssozialismus*. Berlin.

Wagner, A. 1899. *Finanzwissenschaft.* Leipzig.

Wagner, A. 1912a. *Die Strömungen in der Sozialpolitik und der Katheder- und Staatssozialismus*. Berlin: Volkstümliche Bücherei.

Wagner, A.D.H. 1912b. *Finanzwissenschaft, Britische Besteuerung im 19. Jahrhundert und bis zur Gegenwart (1815–1910)*, vol. 2(3), 2nd ed. Leipzig.

Weber, M. 1922. *Roscher und Knies und die logischen Probleme der historischen Nationalökonomie, gesammelte Aufsätze zur Wissenschaftslehre.* Tübingen: Mohr.

German Hyperinflation

Theo Balderston

Abstract

The erratic development of inflation in Germany during the First World War into the hyperinflation of 1922–3 has served as a major test-bed of monetary theory ever since.

This article charts contemporary and modern explanations of its genesis, stabilization and effects. Modern analysis focuses on the interaction between fiscal and accommodating monetary policy and the expectations of financial asset holders; it disagrees, as did contemporaries, over the degree of agency of the government in determining its own deficit. After an optimistic 'Keynesian' assessment of its effects in growth, more recent scholarship has relapsed into pessimism as to its effects on investment.

Keywords

Bubble; German hyperinflation; Germany, economics in the 20th century; Gold standard; Hyperinflation; Inflation expectations; Labour hoarding; Money supply; Public debt; Quantity theory of money; Rational expectations; Seigniorage; Velocity of circulation

JEL Classifications

F; E31; E63; N14

German hyperinflation after the First World War originated in the decision of July/ August 1914 to suspend the gold convertibility of the mark and associated gold- reserve requirements. As with other hyperinflations, this one was irregular. German wholesale prices slightly more than doubled during the First World War. By February 1920 the ratio to 1913 prices was about 17, but then fell, irregularly, to a ratio of 13 in May 1921. After May 1921 inflation resumed and between then and June 1922 average monthly inflation was 13.5 per cent; in the following 12 months it reached 60 per cent (including a short cessation in early 1923 as the Reichsbank temporarily pegged the exchange rate), and 32,700 per cent or about 20 per cent per day between June and November 1923. The mark was stabilized in later November 1923 at one million millionth of its 1913 dollar exchange rate. Although only the period from June 1922 was 'hyperinflationary' (above 50 per cent per month), this period cannot be studied independently of the preceding inflationary history (Holtfrerich 1986).

Contemporary explanation was highly politicized (Kindleberger 1984a). The 'quantity theory' was adopted, especially by the French, to prove the agency of the German authorities in causing the inflation, allegedly in order to undermine the reparations regime. The official German counter-explanation was a variant of the 'quantity theory' known as the 'balance of payments' theory, whereby a budget deficit and its monetization followed inexorably from the exchange-rate collapse, which they blamed on the Treaty of Versailles and its reparations demands (see Williams 1922). The quantity theory presumed a constant velocity of circulation, which was at variance with the facts (Graham 1930; Bresciani-Turroni 1931); an intellectually satisfying resolution of this puzzle awaited Cagan's (1956) embodiment of 'expected inflation' as an argument in the demand-for-money function. The rational expectations' revolution, however, argued that Cagan's formulation of price expectations as a weighted average of past inflation was rational only if the money supply were endogenously determined (Sargent and Wallace 1973).

The question whether German hyperinflation was a 'bubble' divorced from monetary 'fundamentals' continues to be discussed, but the evidence remains inconclusive (for example, Chan et al. 2003). The centrality of fiscal policy and seigniorage to the generation of the German hyperinflation is generally agreed. It is the starting point of Webb's (1989) analysis. The Reichsbank, considering Germany still effectively in a state of war, subordinated its monetary policy to the financing of the Reich's expenditure. Though scarcely stable, a real deficit persisted throughout the inflation, albeit with some tendency to decline as inflation accelerated. The private sector's real investment in debt diminished as its belief weakened in the sufficiency of future budget surpluses to meet the state's contractual debt-servicing obligations (including reparations). The private sector inferred from this insufficiency that prices would rise to reduce the real value of this debt- servicing, and converted its non-monetary debt into money and money into goods. This forced greater monetization of the budget deficit; and the conjuncture of the declining real demand for money with

rising nominal supply made the public expectation of inflation self-realizing. 'Unpleasant monetarist arithmetic' would probably have produced an analogous result even with Reichsbank independence (Holtfrerich 1986, pp. 172ff.).

Frenkel (1977) sought direct evidence of inflationary expectations from the forward discount on the mark in the London foreign exchange market; but, awkwardly from an analytical point of view, until July 1922 the mark sold at a forward *premium*. Webb argued that this reflected the animal spirits of – mainly foreign – speculators with their diversified portfolios, rather than inflationary expectations; these he inferred from the rate of shrinkage of the real value of government debt. On this basis he could link the major shifts in the rate of inflation with announcements of fiscal 'news' that prompted state debt-holders into revising their previous estimates of future real budget surpluses. Plausible connexions of this sort can be made for November 1918 (the Armistice), May 1919 (publication of the Treaty of Versailles), May 1921 (announcement of the Allies' London Reparations Plan) and June 1922 (refusal of a bankers' committee headed by J.P. Morgan Jr. to recommend a loan to Germany except on the – at that point unlikely – condition of a reduction in Allied reparations claims).

Webb explained the sudden cessation of inflation in March 1920 by a conjectural calculation that the expected revenues from the new federal direct taxation introduced by Finance Minister M. Erzberger in 1919 now harmonized with debt obligations (though the reparations obligation was still undefined). He explained the stabilization in November 1923 with reference to the cessation, in late September, of state-subsidized 'passive resistance' against the Franco-Belgian occupation of the Ruhr; to the imposition of indexed tax liabilities from October (see Franco 1990); to the appointment of the Dawes Committee to propose a temporary rescheduling of reparations; possibly to awareness that the Reichsbank was at last threatening to use the independence granted it in May 1922 to cease monetizing the deficit from the end of 1923; and to the successful pegging of the exchange rate against the dollar in mid-November. These developments have to be assumed to have influenced the minds of state debt-holders more than the evidence of the disintegration of the Reich, the collapse of the majority coalition on 3 November, and the lack of clarity, in the hour of France's triumph, over what level of reparations' revision would actually be agreed. Perhaps, after the trauma of hyperinflation, the 'credibility bar' over which stabilization policy had to jump was much lowered (Horsman 1988, p. 33).

The 'Structural School' (Kindleberger 1984b; see Alesina and Drazen 1991) argues that domestic social conflict, especially on the labour market and partly operating through non-budgetary channels, was central to the hyperinflation. Burdekin and Burkitt (1996) focus on the hugely increased discounting of private- sector bills at the Reichsbank from mid-1922, in order (in their view) to pre-finance inflationary wage settlements. Prior to this, foreign speculation in the mark had financed bank lending to business at negative real rates of interest, so that domestic distributional conflicts could be assuaged out of the wealth of foreigners (Holtfrerich 1986, pp. 279ff.). However, once the forward exchange rate flipped over to discount in July 1922, in the absence of Reichsbank accommodation business would have had to pay positive real interest rates, with a correspondingly deflationary effect.

Webb (1989, p. 42) denied that inflation was deliberate government policy. The only reason that the stabilization after March 1920 did not 'stick' was that the Allies' 'London Plan' of May 1921 derailed it; without this element, the Erzberger fiscal reforms were propelling the budget towards surplus. It was irrational to operate in a hyperinflationary zone when, according to the theoretical consensus, real seigniorage revenues would have been greater at a lower rate of inflation. Webb also accepted the 'structural' case that parliamentary conditions and civil-service wage pressure prevented further fiscal reform before autumn 1923 (see Kunz 1986). Cukierman (1988), however, argued for government agency in the inflationary process on the grounds that, due to increasing lags of inflationary expectations behind actual inflation, it could temporarily increase its seigniorage by increasing inflation,

even if at the expense of lower seigniorage in the longer run. The foreshortened time preference of the Reich during its acute diplomatic crisis with the Allies made this rational. Only when expected inflation entered the zone where seigniorage revenues declined – partly due to substitution of other currencies (Bernholz 1995) – did the government stabilize. Cukierman combines this with an argument that the government and the electorate in any case preferred lower long-run seigniorage revenues as these curbed the reparations rapacity of the Allies.

Holtfrerich (1986, pp. 203–05) argued that the inflation counterfactually raised output by neutralizing the effects of the global post-war slump, and equalized income and wealth (but see Kindleberger 1994). However, the ultra-low unemployment of the period was also partly due to vast labour hoarding by public enterprises, dating from the demobilization, and to a trough in participation rates. Bresciani-Turroni (1931, pp. 197–203, 403) argued that the inflation caused misallocation of investment; but Holtfrerich argued (1986, pp. 205–06) that not this misallocation but the deflationary gold-standard regime from 1924 caused the low-capacity utilization of the later 1920s. However, Lindenlaub's (1985) archival investigation concluded that, except for industries receiving government compensation for treaty losses, real fixed investment was minimal (see Fischer et al. 2002).

Sargent (1986, pp. 40ff.) argued that the credibility of the German stabilization made it virtually costless. But Dornbusch (1987) regarded the willingness to make monetary policy hurt from November 1923 to June 1924 as necessary to establishing credibility. The 'stabilization boom' of the second half of 1924 and the delayed but sharp year-long recession from June 1925 may roughly replicate recent high- inflationary experience (Fischer et al. 2002).

See Also

- ▶ Germany, Economics in (20th Century)
- ▶ Hyperinflation
- ▶ Inflation Expectations
- ▶ Quantity Theory of Money
- ▶ Rational Expectations

Bibliography

Alesina, A., and A. Drazen. 1991. Why are stabilizations delayed? *American Economic Review* 81: 1170–1188.

Bernholz, P. 1995. Currency competition, inflation, Gresham's law and exchange rate. In *Great inflations of the 20th century. Theories, polices, evidence*, ed. P. Siklos. Aldershot: Edward Elgar.

Bresciani-Turroni, C. 1931. The economics of inflation: A study of currency depreciation in post-War Germany. London: Allen & Unwin, 1937.

Burdekin, R.C.K., and P. Burkitt. 1996. *Distributional conflict and inflation: Theoretical and historical perspectives*. Basingstoke: Macmillan.

Cagan, P. 1956. The monetary dynamics of hyperinflation. In *Studies in the quantity theory of money*, ed. M. Friedman. Chicago: University of Chicago Press.

Chan, H.L., S.K. Lee, and K.-Y. Woo. 2003. An empirical investigation of price and exchange-rate bubbles during the interwar European hyperinflations. *International Review of Economics* 12: 327–344.

Cukierman, A. 1988. Rapid inflation: Deliberate policy or miscalculation? *Carnegie- Rochester Series on Public Policy* 29: 11–76.

Dornbusch, R. 1987. Lessons from the German inflation experience of the 1920s. In *Macroeconomics and finance. Essays in honour of Franco Modigliani*, ed. R. Dornbusch, S. Fischer, and J. Bossons. Cambridge, MA: MIT Press.

Fischer, S., R. Sahay, and C.A. Végh. 2002. Modern hyper- and high inflations. *Journal of Economic Literature* 40: 837–880.

Franco, G.H.B. 1990. Fiscal reforms and stabilisation: Four hyperinflation cases examined. *Economic Journal* 100: 176–187.

Frenkel, J.A. 1977. The forward exchange rate, expectations, and the demand for money: The German hyperinflation. *American Economic Review* 67: 653–670.

Graham, F.D. 1930. *Exchange, prices, and production in hyper-inflation: Germany 1920–23*. Princeton: Princeton University Press.

Holtfrerich, C.-L. 1986. *The German inflation 1924–23: Causes and effects in international perspective*. Berlin: De Gruyter. German original, 1980.

Horsman, G. 1988. *Inflation in the twentieth century: Evidence from Europe and North America*. Hemel Hempstead: Harvester Wheatsheaf.

Kindleberger, C.P. 1984a. *A financial history of Western Europe*. London: Allen & Unwin.

Kindleberger, C.P. 1984b. A structural view of the German inflation. In *The experience of inflation*, ed. G.-D. Feldman, C.-L. Holtfreich, and G.A. Ritter. Berlin: De Gruyter.

Kindleberger, C.P. 1994. Review: The great disorder: A review of the book of that title by Gerald D. Feldman. *Journal of Economic Literature* 32: 1216–1225.
Kunz, A. 1986. *Civil servants and the politics of inflation in Germany 1914–1924*. Berlin: De Gruyter.
Lindenlaub, D. 1985. *Maschinenbauunternehmen in der deutschen Inflation*. Berlin: De Gruyter.
Sargent, T.J. 1986. *Rational expectations and inflation*. New York: Harper & Row.
Sargent, T.J., and N. Wallace. 1973. Rational expectations and the dynamics of hyperinflation. *International Economic Review* 14: 328–350.
Webb, S.B. 1989. *Hyperinflation and stabilization in Weimar Germany*. New York: Oxford University Press.
Williams, J.H. 1922. German foreign trade and reparations payments. *Quarterly Journal of Economics* 36: 482–503.

German Reunification, Economics Of

Jennifer Hunt

Abstract

German reunification in 1990 posed the challenge of introducing markets to an economy with none. For citizens of the formerly Communist East Germany, the transition brought an immediate increase in political freedom and living standards, yet also a deep trough in output and persistent unemployment. I examine the reasons for the output trough and the subsequent labour market difficulties, analyse the impact of reunification on West Germany and Europe, and draw lessons for transition and economics generally.

Keywords

'Big bang' reform; Active labour-market programmes: in Germany; Communism; Czech Republic; European Monetary Union (EMU); Exchange Rate Mechanism (ERM) (EU); German monetary union; German reunification, economics of; Investment subsidies: in Germany; Labour productivity: in Germany; Migration and labour markets; Privatization: in Germany; Trade unions: in Germany; Transition; Unemployment insurance: in Germany; Unemployment: in Germany; Wage floor theory; Welfare state: in Germany

JEL Classifications

O2; E6; J4; J6; P3; F3; P2

On 3 October 1990 the formerly Communist German Democratic Republic joined the Federal Republic of Germany, thereby reunifying Germany and posing the challenge of introducing markets to an economy with none. German reunification was part of the dramatic demise of Communism in Europe, an event as significant for economic as for political reasons. For citizens of the former German Democratic Republic (henceforth East Germany), the transition brought an immediate increase in both political freedom and living standards, yet also a large rise in economic uncertainty, manifested not least through the sudden emergence of high unemployment. Although markets and institutions were successfully introduced, they have not led to the rapid economic convergence of the two parts of Germany for which some had hoped, and unemployment has remained high. The enormous costs of reunification have proved a burden for West Germany, which prior to unification had been the economic engine of Europe. The shock of unification and the subsequent slow growth in West Germany have in turn affected the rest of Europe.

Historical and contemporary factors ought to have ensured the best outcomes of any transition economy. Before the Second World War, East German GDP per capita was slightly above the German average (Sinn and Sinn 1992), and both at that time and under Communism East Germany was richer than (other) eastern European countries. East Germany's relatively small population – 20 per cent of unified Germany – made feasible the large financial transfers from its rich cousin, West Germany. East Germany has benefited from West German institutions, know-how and investment. Yet the Czech Republic had a GDP per capita only 13 per cent lower than that of East Germany in 2004 (OECD 2005), and, if post-1999

trends continue, the Czech Republic will converge with West Germany before East Germany does.

In this article, I note the successful introduction to East Germany of markets, institutions, democracy and rule of law, and assess why the short-term cost in terms of output and employment was so high. I examine the reasons for the subsequent labour market difficulties, analyse the impact of reunification on West Germany and Europe, and draw lessons for transition and economics generally.

Chronology of Unification

The process culminating in the unification of Germany was set in motion when the Hungarian government began allowing East German citizens to leave Hungary via Austria in May 1989. This occurred against the backdrop of reforms in the Soviet Union by President Michael Gorbachev. By August, large numbers of East Germans were reaching West Germany via Hungary, Czechoslovakia and Poland, and in September anti-government demonstrations began in East German cities. On the night of 9 November 1989, a combination of government weakness and confusion led to a crowd being permitted to breach the wall dividing Berlin. The ensuing mass migration to the West removed the power of the East German government to threaten its citizens: five per cent of the eastern population emigrated in 1989–1990.

The East German government organized free elections for March 1990. The victory of the counterpart of the western Christian Democrat Party was seen as a mandate for rapid reunification. Monetary, economic and social union occurred on 1 July 1990. Political union followed on 3 October 1990. As East Germany was formally joining the Federal Republic of Germany, all western institutions were transferred, and only a small number were subject to a transition period. The western systems of justice, regulation, industrial relations, banking, education and social security and welfare were all transplanted, to a large degree by experts from the west.

Faced with the task of integrating a region with decrepit infrastructure, outdated technology and no capitalist experience, the West German government confronted a number of important decisions in 1990. These included: the exchange rate at which to effect monetary union; how to privatize eastern firms; how to spend money in the east, especially how to spend on consumption versus investment (and infrastructure) and on capital versus labour, and the amounts and details of these expenditures; and whether to raise the money through taxes or debt. Important early decisions by other actors included the decision of labour unions to follow a high-wage strategy.

The financial implications of the government's decisions were colossal. From 1991 to 2003 the west spent four to five per cent of its GDP yearly on the east, including transfers within the social welfare system (Ragnitz 2000, and updated numbers provided by Ragnitz to the author). This spending represented more than 50 per cent of eastern GDP in 1991, and stabilized at about 33 per cent in 1995.

Economic Progress of East Germany

Table 1 documents the evolution of various indicators in east and west. Reunification precipitated a disastrous collapse in real eastern GDP, with falls of 15.6 per cent in 1990 and 22.7 per cent in 1991, cumulating to a one-third decline. Meanwhile, West Germany experienced two boom years with growth rates of over five per cent. From 1992, East Germany experienced four years of recovery followed by stagnation. Growth in the west has also been lacklustre since 1992.

Labour productivity growth in the east was very rapid through 1994, but has since been modest, although higher than in the west. The eastern capital stock, on the other hand, grew at almost six per cent per year or more through 1998, and has continued to grow faster than the western stock since then. Emigration and a plunge in fertility (a 54 per cent fall between 1988 and the 1994 trough) have caused the eastern population to decline each year since unification. Meanwhile, the western population grew quickly in 1990–2

German Reunification, Economics Of, Table 1 Percent change in real GDP, productivity, capital and population, 1990–2004

Year	GDP		Productivity		Capital stock		Population	
	East	West	East	West	East	West	East	West
1990	**–15.6**	**5.7**	–	–	–	–	**–2.5**	**1.6**
1991	**–2.7**	**5.1**	–	–	–	–	**–1.5**	**1.2**
1992	6.2	1.7	18.3	0.7	6.3	2.9	**–0.7**	**1.2**
1993	8.7	–2.6	11.0	–1.4	7.1	2.5	**–0.6**	**0.7**
1994	8.1	1.4	6.3	2.0	7.4	2.1	**–0.4**	**0.4**
1995	3.5	1.4	2.1	1.5	7.4	2.0	**–0.4**	**0.5**
1996	1.6	0.6	2.6	0.7	6.8	1.8	**–0.3**	**0.4**
1997	0.5	1.5	1.9	1.4	6.5	1.7	**–0.4**	**0.2**
1998	0.2	2.3	0.1	0.9	5.8	1.7	**–0.5**	**0.1**
1999	1.8	2.1	1.4	0.7	4.9	1.9	**–0.5**	**0.3**
2000	1.2	3.1	1.7	0.8	4.3	2.0	**–0.6**	**0.3**
2001	–0.6	1.1	0.6	0.3	3.7	1.9	–0.6	0.4
2002	0.2	0.1	1.8	0.4	2.6	1.7	–0.7	0.4
2003	–0.3	–0.1	0.9	0.9	–	–	–0.6	0.2
2004	1.3	1.6	1.0	1.2	–	–	–0.6	0.1

Sources: GDP, productivity, capital stock, population from 2001: Statistische Ämter des Bundes und der Länder. GDP growth in 1990, 1991: Burda and Hunt (2001). Population until 2000: Statistisches Bundesamt.
Notes: Berlin is included in the eastern statistics, except for the figures in boldface, where east and west Berlin are included in eastern and western statistics respectively. West Berlin is about 13% of the population of the 'greater east'. Productivity is measured as GDP per worker. The change in the eastern population in 1989 was –2.5%.

with the arrival of East Germans and immigrants from ex-Communist countries other than East Germany.

Table 2 represents key indicators as the ratio of east to west. Eastern GDP per capita improved from 49 per cent of the western level in 1991 to 66 per cent in 1995, since when convergence has stalled. Because many of the transfers from the west have been to consumption, disposable income per capita has reached a considerably higher plateau, at 81–3 per cent. Capital per worker has continued to converge gradually where other measures have stalled, reaching 84 per cent of the western level in 2002. Compensation per worker rose rapidly from 34 per cent in 1990 to 56 per cent in 1991 and 68 per cent in 1992, and then stabilized at 79 per cent in 1995.

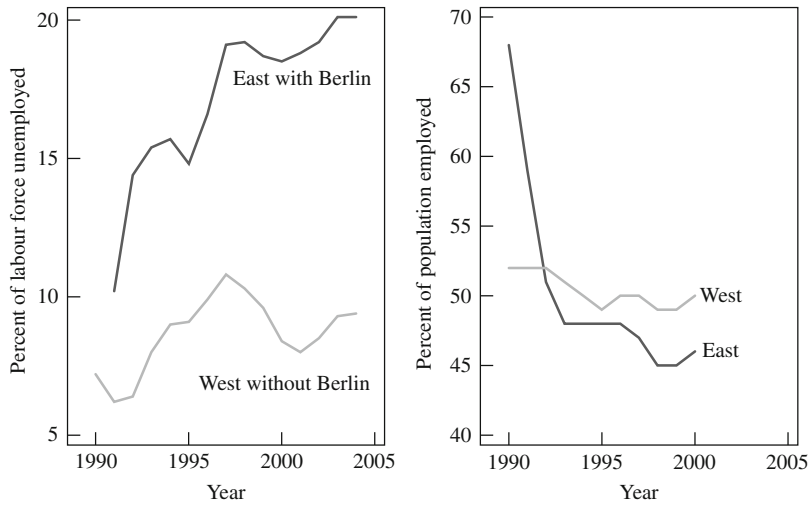
Reunification might be considered a success in terms of standard of living were it not for problems in the labour market. The left panel of Fig. 1 shows the share of the labour force registered as unemployed soared to 20 per cent (from officially zero at the start of 1990), while the western rate has also ratcheted up to a higher level than in 1990. The lack of a search requirement for registering as unemployed means these rates are overstated by several percentage points. The eastern rate is nevertheless very high, especially as some of the many active labour market programme participants would have been unemployed had they not been in the programme. The German Socio-Economic Panel data for the mid-1990s indicate that 15 per cent of the eastern female population and ten per cent of the male population were unemployed (searching and available). The right panel of Fig. 1 shows the plunge in the eastern employment rate.

Splitting the east into its constituent states changes the picture little. The unemployment rates differ little across the six federal states of East Germany. Furthermore, with the exception of unified Berlin, the differences in GDP per capita across the six eastern states are small compared with the east–west gap. This may be seen in Fig. 2, which plots real GDP per capita for Lower Saxony (Niedersachsen), the poorest western state in

German Reunification, Economics Of, Table 2 Measures of convergence – East–West ratios, 1990–2004

Year	GDP			Disposable income	Capital	Compensation	
	per capita	per worker	per hour	per capita	per worker	per worker	per hour
1990	–	–	–	–	–	34	–
1991	49	51	–	63	47	56	–
1992	53	60	–	67	54	68	–
1993	60	68	–	74	57	74	–
1994	64	70	–	77	59	77	–
1995	66	71	–	81	61	79	–
1996	67	72	–	83	64	80	–
1997	67	73	–	83	68	80	–
1998	66	72	67	82	72	81	74
1999	66	72	69	83	75	81	74
2000	66	73	68	82	79	81	75
2001	65	73	68	81	82	81	76
2002	66	74	71	82	84	82	76
2003	67	74	71	82	–	82	73
2004	67	74	–	–	–	82	–

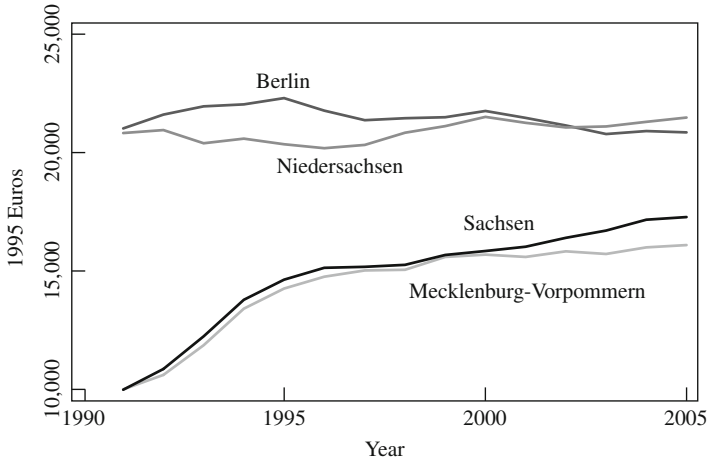
Sources: Statistische Ämter des Bundes und der Länder; author's calculations. For 1990, German Institute for Economic Research, Berlin; data on GDP, employment and compensation in East Germany (without West Berlin) from 1989 to 1998 no longer available on the Institute's website.
Notes: East as a percent of west. Berlin is included in the eastern statistics, except for the figures in boldface, where east and west Berlin are included in eastern and western statistics respectively. 1990 figures are for the first quarter, seasonally adjusted. Productivity is measured as GDP per worker or GDP per hour worked.



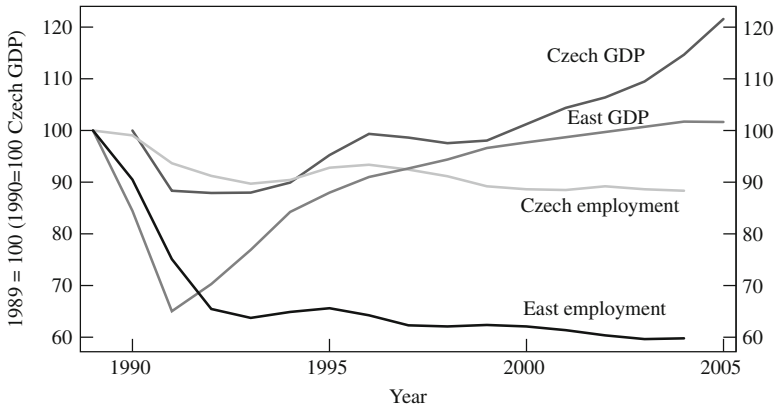
German Reunification, Economics Of, Fig. 1 Unemployment and employment rates, 1990–2004

2004, Saxony (Sachsen), the richest eastern state in 2004, and Mecklenburg–Vorpommern, the poorest eastern state in 2004.

Even the fastest-growing states of Sachsen and Sachsen–Anhalt are growing considerably more slowly than the Czech Republic, as may be seen with the aid of Fig. 3. While East German employment languishes at 60 per cent of its 1989 level, and real GDP has barely risen above its 1989 level, Czech GDP is 20 per cent above its 1990



German Reunification, Economics Of, Fig. 2 Real GDP per capita for selected states



German Reunification, Economics Of, Fig. 3 Czech and East German comparisons

level, and, while Czech employment has not recovered from liberalization, it fell much less than East German employment.

Explaining the Initial Collapse of GDP and Employment

All former Communist countries except China experienced output declines following price liberalization, and many countries of the former Soviet Union had larger and longer output falls than East Germany. Roland (2000) examines why price liberalization depressed output, emphasizing theories of disruption of supply chains and the need to identify new business partners before investing. The main other potential culprits for the GDP and employment declines in East Germany are a reduction in labour supply, substitution to western products, the exchange rate chosen for monetary union, the increase in wages, and the privatization process.

Reduction in Labour Supply

Some of the output decline could have been caused by the employment decline rather than the reverse. Employment declined by 3.3 million people from 1989 to 1992. The government paid one million people to stop working by offering early retirement onto the western pension benefits implied by easterners' years of work experience. A further one million people emigrated to the west

in 1989–91 (Hunt 2006, draws lessons from eastern emigration).

Among the prime-aged remaining in the east, women experienced a particularly large employment decline, a fact often explained by the dismantling of the Communist day-care system. However, Hunt (2002) shows that the employment rate of women with small children fell by no more than that of other women.

Substitution to Western Goods

Immediately upon monetary union, eastern shops filled with western goods. Easterners wanted to consume western products, and at this time 'one couldn't sell an East German egg' (personal communication from eastern state politician; see also Sinn and Sinn 1992). Economists agree that this caused a sudden fall in demand for eastern goods, and hence a fall in output.

Monetary Union Exchange Rate

For political reasons, the (western) government decided to choose a one-to-one exchange rate between the eastern Ostmarks and the western Deutschmarks. Early studies, in particular, argued that an overvalued exchange rate had made the eastern products uncompetitive with western products, leading to an output decline. With hindsight, it seems unlikely that the exchange rate was an important contributor to the output decline, as eastern prices and wages subsequently rose, rather than falling to correct the real exchange rate.

Unions and the Wage Increase

Although it is possible that the rapid rise in wages was the result of factor price equalization across regions, there is a consensus that labour unions were the driving force behind the rise. Unions acquired great power at a time when employers had little, and were not acting only in the interests of eastern workers. Western unions established themselves in the east in 1990, and were very successful in recruiting new members. The new eastern unions were led by westerners, who were concerned with east–west equity and eastern welfare but also with western wages and the perceived threat to them of mass east–west migration. The unions pushed for rapid wage convergence with the west, believing this was just, would prevent mass migration, and would enable eastern workers who were laid off to receive higher unemployment benefits (these being tied to the pre-layoff wage). At this time, most firms had no owners, and the unions were bargaining either with managers, who had no incentive to resist wage increases, or even with members of the western employers' federations, whose incentive was to prevent undercutting of western prices by eastern firms.

Most economists believe this rapid rise in wages represented a classic textbook wage floor that reduced employment, led to high unemployment, and made East German firms unviable, thereby leading to the output collapse (for example, Akerlof et al. 1991; Sinn and Sinn 1992; Sachverständigenrat 2004).

Privatization

Small, mostly service firms were privatized separately from large industrial firms. As in eastern European countries, this privatization was rapid and successful, and was completed by March 1992 (Sinn and Sinn 1992). Large industrial firms were privatized by a politically independent body known as the Treuhandanstalt (THA). Its initial portfolio was 8,500 previously state-owned enterprises containing 44,000 plants and 45 per cent of the workforce (Carlin 1994).

The THA closed the unviable firms and plants, reduced employment at the viable plants, and sought buyers for the remaining core businesses. The THA's aim, at which it was successful, was to match firms with western management expertise in the same industry (Dyck 1997). Weighted by employment, 74 per cent of sales were to West German firms or families, six per cent were to non-German firms, and only 20 per cent were to eastern buyers. Privatization thus created subsidiaries of western companies (Carlin 1994). By 31 December 1994, the THA had finished its privatization with net losses of DM 193 billion (about 95 billion euros or 120 billion US dollars; Brada 1996).

The THA destroyed many jobs in the short run, with the aim of curtailing inefficient production and promoting faster medium-run employment

and output growth than would otherwise have occurred. Most economists studying privatization believe that the THA carried out its mandate well, leaving a legacy of viable and well-run companies. However, Roland (2000) believes that the employment reduction necessitated by the mandate caused a depression in the short run and retarded transition in the medium run.

Explaining the Persistent Labour Market Problems

Even observers who did not expect faster GDP convergence than has occurred are dismayed at the state of the labour market. Most explanations for the initial employment collapse apply to the short run only. Even labour union power has been severely weakened: while unions controlled wages from 1990 to 1993, a subsequent employer revolt allowed wages to be determined more freely. The share of workers whose employer belonged to an employer federation, which determines whether workers are paid the union wage, declined from 76 per cent in 1993 to 45 per cent in 1998 and 29 per cent in 2003 (Brenke 2004).

Either the causes of the initial collapse have had lasting effects – for example, perhaps it is hard to reduce real wages in a low-inflation environment – or there must be other explanations. The leading one is the introduction of the western social welfare system. Others are investment subsidies, the wholesale transfer of western regulations, ineffectual active labour market programmes and impediments to the optimal allocation of resources across sectors.

Social Welfare and Wage Floors

Many economists (for example, von Hagen and Strauch 1999) stress the disincentives of the social welfare system as a cause of low employment in East Germany. After a very brief transition period, benefits were set at western levels, which in some cases made them higher relative to wages than in the west. This was the case in particular with *Sozialhilfe,* or social assistance (welfare), and with pensions. Unemployment insurance benefits are a fraction of the previous wage, so, to the extent that unemployment insurance is a greater problem than in the west, it is related to wages being too high. A generous social safety net sets a floor under wages, similar to a union wage, though affecting labour supply rather than labour demand.

The wage floor theory implies that wages at the bottom of the distribution should have risen the most, while employment of the least skilled should have fallen the most. Employment rates indeed fell more for the less skilled than the skilled. However, wage growth for the skilled was equal to or greater than that of the unskilled (Burda and Hunt 2001). Furthermore, by 1999 wage inequality and the wage structure more generally were very similar to those in the west. Patterns of unemployment duration were also similar (Hunt 2004). These results are inconsistent with the effect of a wage floor for the less skilled, which appears to rule out the social welfare theory. However, it is possible that a wage floor was too simple a model for the effects of the unions, who indeed appeared to aim to raise the wages of all members.

Investment Subsidies

At least with hindsight, subsidizing capital (investment) in the face of grave labour market difficulties seems not obviously a good idea. Indeed, the capital–labour ratio in manufacturing is now higher in the east than the west (Sachverständigenrat 2004). Furthermore, many have criticized the subsidies as being skewed towards structures at the expense of equipment (for example, Burda and Hunt 2001). Finally, the subsidies were designed as tax breaks, and were hence attractive only to profitable, that is, established western, companies. The funds for investment subsidies appear not to have been spent optimally.

Active Labour Market Programmes

Easterners are well educated, and the return to eastern schooling was not reduced by transition (Krueger and Pischke 1995). The post-unification fall in the return to experience indicated that the human capital lacking was experience working in capitalist firms. Off-the-job training and make-

work jobs were therefore unlikely to be very helpful, despite the large number of participants: in 1994 there were 259,000 participants in public training programmes and 280,000 participants in jobs whose wage was paid by the government, compared with 1,142,000 registered unemployed.

The best-documented effect of training programmes has been that of keeping participants out of the labour force for the duration of the sometimes long programmes (Lechner et al. 2007). Meanwhile, participants in public jobs had no incentive to look for another job, as they received 100 per cent of the union wage (90 per cent from 1994 on). While some groups have benefited from some public programmes, the gains are unlikely to have justified the large expenditures (Eichler and Lechner 2002, (Lechner et al. 2007).

Sectoral Allocation

Various factors may have intervened to prevent an optimal allocation of resources across sectors. Brada (1996) observes that the THA requirement that buyers continue operating the firm in the same industry as before may have delayed sectoral restructuring. Unions, bargaining at the industry level, may have chosen the wrong wage structure across sectors, reducing incentives for restructuring (Burda and Hunt 2001; Hunt 2001). A further complicating factor has been the boom and subsequent bust of the construction industry. Many observers believe the manufacturing sector is too small, at 15 per cent of employment in 2004 compared with 22 per cent in West Germany and 30 per cent in the Czech Republic. Yet manufacturing in the United States employed a smaller share of the workforce than in East Germany in 2004, so East Germany may simply have leapfrogged West Germany in this regard.

Effect of Unification on the West

In the short term, reunification was a positive aggregate demand shock for the west, leading to the boom seen in Table 1. The leap in the demand for capital for investment in the east, combined with the reduction in the money supply to contain inflation, raised the interest rate. As the cost of reunification became clear, the government was forced to raise taxes, but debt rose from 41.8 per cent of GDP in 1989 to 64.2 per cent in 2003. The budget went from surplus in 1989 to a 3.1 per cent deficit in 1991, and has been close to or above three per cent since then.

It is unclear to what degree the western stagnation that has followed the 1993 end of the boom can be attributed to reunification. While exports have recovered, domestic demand has remained weak (Sachverständigenrat 2005). This could possibly be the result of government debt leading consumers to revise their wealth downwards, depressing consumption and growth (Carlin and Soskice 2006). The increase in western unemployment, seen in Fig. 1, could be caused in part by increases in payroll taxes to finance reunification. On the other hand, Siebert (2005) emphasizes that before reunification West Germany had already had problems with sluggish growth, rising unemployment, and funding social security.

Posen (2005) considers that approximately 1.4 per cent of German GDP per year is paid in transfers to the east that are for neither investment nor infrastructure, nor part of the unified social welfare system. He calculates the opportunity cost of this money (that could have been invested and received a return), the increase in interest payments on other debt (owing to a higher interest rate caused by higher debt), and the deadweight loss from increased taxes. He concludes that the burden of these transfers is (at most) 0.7 per cent of German GDP per year, a large sum.

Reunification has affected the West German labour market through the weakening of labour unions caused by the collapse of eastern unions. The impact of eastern immigrants and commuters is not known. The impact, if any, would have been in addition to that of the concomitant and similarly sized immigration of ethnic Germans from other formerly Communist countries.

Effect of Unification on Europe

The rise in the German interest rate had important consequences for Europe, as it led to a crisis in the

European Exchange Rate Mechanism (ERM) that preceded European Monetary Union. The higher German interest rate meant that the Deutschmark required a revaluation within the ERM, or, equivalently, the devaluation of other ERM currencies. France and other countries attempted to maintain the existing exchange rates, fearful of a loss of deflationary credibility. But in 1992 speculative attacks forced several countries to devalue, while the United Kingdom and Italy left the ERM.

The crisis was not all bad in the long run: for the United Kingdom, which had joined the ERM at an unsuitable exchange rate, leaving the ERM proved to be an economic boon (Carlin and Soskice 2006). However, Germany may have entered monetary union at a rate that would prove overvalued once the reunification shock to interest rates had passed (Sinn 1999), thus requiring a later depreciation. The difficulty of price and wage adjustments within monetary union may currently be preventing such a depreciation from occurring, slowing German, and therefore European, growth (Carlin and Soskice 2006).

Lessons Learned

Because East Germany joined the well-functioning and larger Federal Republic of Germany, it could feasibly and credibly have an institutional 'big bang', immediately importing a coherent set of institutions generally suitable for the region. This provided confidence and familiarity to western investors. The institution that obviously made a poor transition was the industrial relations system: because labour unions were established before employer federations, labour unions were initially unnaturally strong, possibly with lasting consequences.

Some economists believe the social welfare system made an equally poor transition; yet the nature of reunification meant that there was politically no alternative to transferring the system fairly rapidly. Siebert (2005) bemoans the transfer of product regulation and taxation. Yet firms may have complied with western constraints even had they not been imposed on the east, either in the expectation of their being imposed later or for fear of disgruntlement in their western works council. For example, Volkswagen applied the western prohibition on female night work to its eastern plant although the east was exempt (Turner 1998).

The feasibility of an institutional big bang made feasible an economic big bang. Price liberalization and macro stabilization were flawless. The privatization process was speedy and had many merits, although it may have led to an excessive employment decline, and was too expensive for most countries to countenance. However, Koreans should note that even an unusual transition that satisfies both the ' Washington consensus' economists, who emphasize speed of economic reform, and the ' evolution-institutionalist' economists, who stress the necessity of establishing institutions before economic reform, can leave in its wake a difficult regional convergence problem.

For economists interested in unemployment, East Germany is both a validation of textbook models and a puzzle. Surely the collapse in employment and output in 1990–2 must have been strongly influenced by high union wages. Yet, now that labour unions have much less influence and the wage structure is similar to that of the west, why has unemployment remained so high? Good education and high emigration are not enough to control unemployment.

See Also

- ▶ Privatization
- ▶ Total Factor Productivity
- ▶ Transition and Institutions
- ▶ Unemployment
- ▶ Unemployment Insurance

Acknowledgment I am very grateful to Michael Burda, Wendy Carlin, Adam Posen and Harald Uhlig for helpful discussions, and to Karl Brenke, Michaela Kreyenfeld, Joachim Ragnitz and Werner Smolny for generously and quickly providing me with data.

Bibliography

Akerlof, G., A. Rose, J. Yellen, and H. Hessenius. 1991. East Germany in from the cold: The economic aftermath of currency union. *Brookings Papers on Economic Activity* 1991(1): 1–87.

Brada, J. 1996. Privatization is transition – or is it? *Journal of Economic Perspectives* 10(2): 67–86.
Brenke, K. 2004. Ostdeutsche industrie: Weitgehende abkehr von der kollektiven lohnfindung. *DIW-Wochenbericht* 13: 5–8.
Burda, M., and J. Hunt. 2001. From reunification to economic integration: Productivity and the labor market in Eastern Germany. *Brookings Papers on Economic Activity* 2001(2): 1–92.
Carlin, W. 1994. Privatization, distribution and economic justice: Efficiency in transition. In *Privatization in Central and Eastern Europe*, ed. S. Estrin. London: Longman.
Carlin, W., and D. Soskice. 2006. *Macroeconomics: Imperfections, institutions and policies*. Oxford: Oxford University Press.
Dyck, I. 1997. Privatization in Eastern Germany: Management selection and economic transition. *American Economic Review* 87: 565–597.
Eichler, M., and M. Lechner. 2002. An evaluation of public employment programmes in the East German state of Sachsen-Anhalt. *Labour Economics* 9: 143–186.
Hunt, J. 2001. Post-unification wage growth in East Germany. *Review of Economics and Statistics* 83: 190–195.
Hunt, J. 2002. The transition in East Germany: When is a ten point fall in the gender wage gap bad news? *Journal of Labor Economics* 20: 148–169.
Hunt, J. 2004. Convergence and determinants of non-employment durations in Eastern and Western Germany. *Journal of Population Economics* 17: 249–266.
Hunt, J. 2006. Staunching emigration from East Germany: Age and the determinants of migration. *Journal of the European Economic Association* 4(5): 1014–1037.
Krueger, A., and J.-S. Pischke. 1995. A comparative analysis of East and West German labor markets: Before and after unification. In *Differences and changes in wage structures*, ed. R. Freeman and L. Katz. Chicago: University of Chicago Press.
Lechner, M., R. Miquel, and C. Wunsch. 2007. The curse and blessing of training the unemployed in a changing economy: The case of East Germany after unification. *German Economic Review* 9(1): 468–905.
OECD (Organisation for Economic Co-operation and Development). 2005. *OECD in figures, 2005*. Paris: OECD.
Posen, A. 2005. Much ado about little: The fiscal impact of German economic unification. *CESifo Forum* 4: 33–36.
Ragnitz, J. 2000. Was kostet die einheit? zur bewertung der transferleistungen für ostdeutschland. In *Nutzen und kosten der wiedervereinigung*, ed. D. Brümmerhof. Baden Baden: Nomos.
Roland, G. 2000. *Transition and Economics: Politics, Markets, and Firms*. Cambridge, MA: MIT Press.
Sachverständigenrat zur Begutachtung der gesamtwirtschaflichen Entwicklung. 2004. *Erfolge im Ausland – Herausforderungen im Inland. Jahresgutachten 2004/05.* Reutlingen: Servicecenter Fachverlage. Online. Available at http://www.sachverstaendigenrat-wirtschaft.de.Accessed 3 May 2006.
Sachverständigenrat zur Begutachtung der gesamtwirtschaflichen Entwicklung. 2005. *Die Chance nutzen – Reformen mutig voranbringen. Jahresgutachten 2005/06.* Reutlingen: Servicecenter Fachverlage. Online. Available at http://www.sachverstaendigenrat-wirtschaft.de.Accessed 3 May 2006.
Siebert, H. 2005. *The German economy.* Princeton: Princeton University Press.
Sinn, G., and H.-W. Sinn. 1992. *Jumpstart: The economic unification of Germany.* Cambridge, MA: MIT Press.
Sinn, H.-W. 1999. International implications of German unification. In *The economics of globalization: Policy perspectives from public economics*, ed. A. Razin and E. Sadka. Cambridge: Cambridge University Press.
Turner, L. 1998. *Fighting for partnership: Labor and politics in unified Germany.* Ithaca: Cornell University Press.
von Hagen, J., and R. Strauch. 1999. Tumbling giant: Germany's experience with the maastricht fiscal criteria. In *From EMS to EMU: 1979 to 1999 and beyond*, ed. D. Cobham and G. Zis. London: Macmillan.

Germany in the Euro Area Crisis

Daniela Schwarzer

Abstract

This article examines Germany's move to centre stage in the management of the sovereign debt crisis that has ravaged the euro area since the beginning of 2010. The German government has been influential in deciding the pace and design of rescue packages, and also in the ongoing reform of governance in the euro area. Germany's dominance can be explained by its large contribution to the rescue packages, by its relative economic strength and by the role of potential veto players such as the German Parliament and the German Constitutional Court. Germany's positions on crisis management and on governance reform reflect its approach during the Maastricht negotiations: to minimise risk-sharing and joint liabilities so as to avoid

moral hazard, to increase control and the credibility of rule-based coordination, to enable European control of national policies, and to strengthen the market mechanism to discipline national policy choices.

Keywords

Angela Merkel; Bundestag; Deutsche Bundesbank; Economic governance; EU governance; Euro; Euro area; Euro zone; Germany; Sovereign debt crisis

JEL Classification

E6; E42; F15; F33; F36; H3; H6; O3; O31; O320; O380

Introduction

Without actively seeking the role, Germany became a key actor in the management of the sovereign debt crisis in the euro area when Greece succumbed to market pressure in early 2010. Thanks to its global competitiveness, Germany displayed comparatively strong resilience to the economic crisis. Due to its relative size, it shoulders the largest single share (27%) of the contributions and guarantees in the rescue mechanisms set up since 2010. This contribution, combined with its singular position of being a large AAA-rated country, enabled it to influence the pace and instruments of crisis management and to shape the subsequent economic governance reforms in the euro area.

This article discusses Germany's approaches to and impact on the management of the sovereign debt crisis and its approach to governance reform in the euro area. It highlights its underlying strategy, Germany's preferences and the determining factors behind them. It also investigates the evolution of Germany's power base and argues that a reassessment of the role of domestic veto players and structural developments in the EU have relativised the view that Germany has become the hegemon in the monetary union.

Germany's Approach to Crisis Management in 2010/2011

Germany's initial reaction to the sovereign debt crisis was one of prudence and caution, being careful not to extend rapid and decisive aid to fellow member states. This became particularly obvious in its successful attempts to decelerate the EU's move towards a first rescue package for Greece. In February 2010, the European Council had promised to help member governments having trouble refinancing their debt (European Council 2010), but only followed through in April 2010. In the eyes of many observers, Germany was delaying the implementation of the rescue package and provoked nervousness in the financial markets. The German government, however, had no desire to actually let Greece fail, given the fact that German banks were the largest holders of Greek government bonds in 2010.

There are several reasons for the German reluctance to provide loans more swiftly. A major concern was to maintain pressure on the Greek government to consolidate its budget and implement structural reforms. Strict conditionality was attached to the loans at Germany's insistence, emphasising the need for both budgetary austerity and structural adjustment in order to reduce moral hazard. Although members of the German government initially argued for a euro-area-only solution to handle the liquidity problems, Chancellor Merkel revised this position in March 2010, when she declared in the German Parliament (*Bundestag*) that loans to illiquid member states would be granted in cooperation with the International Monetary Fund (IMF). Germany also demanded that interest rates on credit given to indebted countries should be high in order to prevent 'easy borrowing', which could impede the implementation of reforms. It further expected compliance with the European rules for fiscal and economic governance as a prerequisite for financial aid. Other governments (for instance Germany's closest partner, France) argued that earlier and more encompassing rescue packages would be crucial to reassure market participants and break the vicious circle of a self-fulfilling crisis.

These positions reveal divergent underlying assumptions about the dynamics of financial crises (Dullien and Schwarzer 2011). Some governments adhered to the idea that financial crises, just like currency crises, could become self-fulfilling prophecies, if only a relevant number of market actors believed there was a likelihood of a crisis; other governments, including Germany's, did not seem to believe that there was a high risk of a rapid crisis escalation, which would be detached from economic fundamentals. Following this logic, there was little need to counter market expectations with financial aid mechanisms that would potentially cover the worst-case scenario – which could only be prevented if market participants thought that there were sufficient tools to achieve this. Therefore, the German government advocated a policy that was designed to improve fiscal balances and real economic situations, while only setting up immediate rescue provisions. Chancellor Angela Merkel consequently only supported the creation of new rescue funds when the crisis escalated to an unexpected degree in May 2010. Given the particularly high exposure of German banks to sovereign debt in Ireland, Italy and Spain at the time, concerns grew in Germany that, if the crisis spread to the two largest Southern European member states, it could entail heavy losses for the German financial sector.

Limiting Spill-Over Through Closer Policy Coordination

In May 2010, the temporary rescue mechanisms – the European Financial Stability Facility (EFSF) and the European Financial Stabilisation Mechanism (EFSM) – were agreed (Council of the European Union 2010). In reaction to Germany's financial involvement, the German debate concentrated on the question of how a similar crisis could be prevented in the future. Germany's share in the financial rescue mechanisms became politically linked to the quest for substantial governance reforms. When convincing the Bundestag to ratify the rescue packages, the government argued that it was working with its European partners to fight the root causes of the crisis. At the core of this debate was how surveillance and policy coordination in the EU could be improved in order to retransform the euro area into a monetary union which reflected the policy and institutional preferences that Germany had sought to enshrine in the Maastricht Treaty (Dyson and Featherstone 1999, pp. 370f).

The consolidation pressure exerted on countries in crisis, coupled with the first German governance reform proposals, clearly demonstrated that the dominant interpretation was that of a fiscal crisis with its root cause in the irresponsible fiscal behaviour of governments. Finance Minister Wolfgang Schäuble tabled a proposal for euro area governance reforms on 21 May 2010 (Bundesfinanzministerium 2010) which included stronger, rule-based fiscal policy coordination involving nominal targets and automatic sanctioning mechanisms, and very little room for political discretion, in order to prevent fiscal policies from undermining monetary stability. In the subsequent debate, proposals went as far as to withdraw voting rights from member states breaking the rules of the Stability and Growth Pact. The idea of a 'Super Commissioner' who would directly control the member states' performance and could interfere with national budgets and hence limit member states' budgetary sovereignty was also discussed. The German concern to improve preventative policy coordination is also reflected in a further German invention, the Fiscal Compact. This intergovernmental treaty was signed on 30 January 2012 by 25 member governments, and obliges its signatories to adopt national fiscal rules as a backup for European coordination mechanisms in order to reduce the risk of budgetary problems and moral hazard in the member states.

The Euro Plus Pact, concluded in March 2011, reflects the concern to push other governments to adopt measures that would increase competitiveness by prescribing a structural reform agenda for the near-future (Council of the European Union 2011a). It partly advocates policies seen as having contributed to Germany's economic success over the last decade and was hence very controversial with member governments that were critical of their emphasis on supply-sided reforms.

Germany's initiatives to push for economic convergence in the euro area were backed by the Federation of German Industry (BDI), which not only underlined the importance of competitive suppliers for Germany's economic success, but also tried to encounter any attempt to push for policies that might lead to a reduction in German price competitiveness.

As concern grew that solvency issues might have to be tackled in the euro area, the German Finance Minister advocated a sovereign default procedure that could be part of a European Monetary Fund (Schäuble 2010). The idea of creating a legal framework and a mechanism for debt restructuring also meant solving the question of the no-bailout clause of Art. 125 in the Lisbon Treaty. When negotiations on the European Stability Mechanism (ESM) advanced, the German Finance Ministry argued that private sector involvement should be institutionalised with it. It even advocated at some point that this should occur quasi-automatically in crises of liquidity – rather than solvency only. Market unrest and resistance from the European Central Bank (ECB) and other governments, however, made German decision-makers reconsider this position. The idea of a mechanism for sovereign debt restructuring did not gain traction. But the ESM Treaty, signed on 2 February 2012, did oblige signatory states to include collective action clauses in newly issued government bonds. Thus, the first elements of a legal framework for debt restructuring in the case of a sovereign default were eventually introduced.

Changing Views on the Crisis

In 2011, the crisis escalated into a new phase. Given unparalleled signals of panic in the markets, the debate re-emerged as to whether the euro area needed a potentially unlimited crisis intervention mechanism to stop expectation-driven crises of confidence. At the time, the idea of introducing euro bonds as a remedy to the prevailing crisis even gained some traction in Germany, supported by the Green Party (*Bündnis 90/Die Grünen*), parts of the Social Democratic Party (*SPD*) and the Confederation of German Trade Unions, but was at no point endorsed by the governing CDU/CSU/FDP coalition.

On 21 July 2011, as a result of a grave deterioration of the situation on the bond markets, the euro area summit decided to expand the credit volume and the instruments of the EFSF (Council of the European Union 2011b). However, there was still no consensus to create a potentially unlimited rescue mechanism. Several governments, including the German one, opposed a substantial increase of the EFSF for Spain and Italy or the introduction of euro bonds, advocated by other member states and partly backed by the European Commission and prominent members of the European Parliament. Rather, they accepted that the ECB intervened heavily in the bond markets. This policy led to an unexpected rift between the government and German central bankers. While Chancellor Merkel and Finance Minister Schäuble silently backed the increasing activities of the ECB to stabilise sovereign debt in the euro area, both the German chief economist of the ECB, Jürgen Stark, and Axel Weber, President of the Bundesbank and a potential candidate for the Presidency of the ECB, resigned from their posts in protest in 2011.

A year later, in August 2012, after further deterioration in the sovereign bond markets, the ECB finally announced that it would be possible to intervene in bond markets in unlimited volumes through the Outright Monetary Transactions (OMT) programme. While the OMT is part of the ECB's monetary policy measures to ensure the transmission of monetary policy decisions, it *de facto* proved to be a cure to the crisis in sovereign debt markets which none of the measures that the euro area governments agreed upon had been able to achieve.

From 2010, most of the attention had been on developments in sovereign debt markets. In late 2012, however, the growing economic and political risks and the social instability in member states that were under particularly severe adaptation pressure demanded increasing attention. New discussions about adequate growth policies evolved, mostly in Southern EU member states, but also in Germany. The 'austerity-first' approach which had dominated the discourse

since 2010 was still present. But a gradual re-think of the approach towards Greece and other crisis countries set in. The German opposition at the time, consisting of the SPD, the Greens and the Left party (*Die Linke*), which was strengthened politically by the election of the Socialist candidate François Hollande as French President in May 2012, caused a delay in the ratification of the fiscal compact and the ESM in the Bundestag. It requested German support for a new European growth strategy and a European Financial Transaction Tax as a precondition for its approval.

Moreover, the strategy to deal with imbalances in the euro area was being cautiously reconsidered. So far, the German position had been that imbalances in countries running an external deficit should be corrected by improving competitiveness and hence generating external growth. Germany had come under pressure from fellow euro area governments, the European Commission and the IMF, which all pointed to Germany's responsibility in the euro area rebalancing process. While Germany had advocated an asymmetric approach in the negotiations leading to the creation of the new instrument to coordinate national economic policies, the Macro-Economic Imbalance Procedure, Finance Minister Schäuble, in May 2012, conceded in an interview that 'it is fine if wages in Germany currently rise faster than in other EU countries. These wage increases also serve to reduce the imbalances within Europe' (Bryant 2012). However, any external policy measures or advice that could be interpreted as undermining German price competitiveness would either be ignored or fought off immediately. The mainstream view in Germany, widely shared by policy makers, the business community and academics, was that rebalancing should be achieved by improving the competitiveness of the deficit countries, and, if at all, a stronger domestic demand in surplus countries like Germany.

Normative Underpinnings of German Preferences

Germany's proposals for economic governance reform in the course of the sovereign debt crisis were contiguous to German positions in the Maastricht negotiations. Germany's idea of the euro area has always been one of a stability union in which an independent, stability-oriented monetary policy would be underpinned by sound fiscal policies and economic policies designed to enhance member states' competitiveness. Policy coordination should be rule-based and should prevent negative spill-overs – rather than ensure the provision of public goods such as growth or employment. Although Germany, over the years, had argued for stronger political integration, and even a political union, the idea was not to introduce political discretion at the euro area level, but rather to strengthen the rules-based coordination approach by enabling the European Commission to interfere more strongly with national policies. This view explains why Germany only hesitantly accepted that a euro area summit – as suggested by the French government – should meet regularly and why it made a concerted effort to pre-define the policy agenda of such a summit (e.g. by laying down detailed policy objectives in the Euro Plus Pact).

The German approach to the functioning of a monetary union tended not to include solidarity or mutual insurance mechanisms. If at all, fiscal solidarity or mutual insurance mechanisms should only exist to a degree that would not undermine the responsibility of national governments to engage in structural reforms and improve competitiveness, but should rather encourage such measures. In 2013, the German government suggested the introduction of a so-called "Convergence and Competitiveness Instrument", also tagged as "Contractual Arrangements", which combined bilateral contracts in which governments would commit themselves to the implementation of certain reforms, and in exchange would receive financial support from a new instrument that still needed to be created. This proposal met with great resistance at the European Council in December 2013, with almost all governments opposing the German initiative. Further discussion was postponed to October 2014. The widespread perception at the time was that Germany was seeking to establish another mechanism by means of which "the German model" could be imposed on others.

Germany was also accused of not showing a sufficient sense of solidarity with the crisis countries. This points to different notions of solidarity prevailing in the euro area at the time. German policy makers would reject any claims that Germany should accept stronger financial transfers – without perceiving a lack of solidarity. For Germany, solidarity is not predominantly interpreted as the readiness to share risks and accept financial transfers, but rather as respecting the rules that should underpin sound public finances, competitiveness and monetary stability, which the member states have agreed upon. The bottom line of Germany's approach to crisis management and governance reform was to limit risk-sharing and to make member states liable for their own debts in the future.

This German understanding of an 'appropriate' functioning of the euro area is often explained by the strong influence of ordoliberalism, a German mid-20th century variant of neoliberalism that drove the creation of the post-Second World War German social market economy. This school of thought argues that central banks should be independent and focused on the provision of monetary stability, while governments should set the framework for functioning markets and free competition.

German views on the euro area and crisis management are also strongly influenced by the paradigm of neo-classical economics, the school of thought that currently dominates teaching and economic research in Germany (see Dullien and Guérot 2012, for an overview). It assumes that (financial) markets set the prices of assets in an adequate manner if they possess sufficient information. Economies are assumed to adjust rapidly to shocks, above all by supply-side reforms instead of the demand-side measures postulated by Keynesian scholars. If demand is not strong enough and the employment situation is deteriorating, a reduction in price and wage levels is recommended to improve competitiveness. In compliance with this conviction, in 2003, the red–green coalition government in Germany implemented 'Agenda 2010' – a package of structural reforms of the German labour market and the social security system. In retrospect, those reforms are widely considered as the precondition of Germany's competitiveness today and its resilience to the sovereign debt and economic crisis. Assuming that the German experience can be copied, many German economists and policymakers have argued that similar structural reforms should be undertaken in countries like Greece, Spain and France – implying that growth and demand will return in the crisis-stricken member states as a result of a successful adaptation strategy.

It is the importance accorded to this supply-side policy that has provoked conflicts between Germany and its European partners. Whereas German policy-makers tend to frame Agenda 2010 as a German success story, some European partners perceive the German supply-side reforms, which lowered the price level, slowed down consumption and balanced economic activity towards exports, as 'beggar thy neighbour' policies. The ambitious structural reforms had a negative impact on demand not only in Germany but also in the countries closely linked to Germany, such as France.

Domestic Politics

The German debate on crisis management was more politicised than in most other member states, partly due to the Bundestag's strong involvement in the decisions on the rescue measures and the intense media coverage thereof. The postponement of a European Council meeting in October 2011 for three days – despite the intense pressure of the crisis – in order to enable the Bundestag to give the Chancellor a mandate, illustrates this new assertiveness by the legislators (Alexander 2011). The coalition government's policies and rhetoric were strongly shaped by parliamentarians' anticipated reactions, potential inner-party conflicts and the strong concern that a Eurosceptic party might eventually take shape. While members of the senior Christian Democrat coalition partner CDU/CSU openly criticised the government's crisis management, the Free Democratic Party (*FDP*) even attempted to stop the ESM through a referendum within the party, but failed to meet the necessary quorum.

Parliamentarians were gravely concerned that the rescue packages, and in particular the establishment of medium- and long-term rescue mechanisms, were compatible with the German constitution. Related questions were whether the Bundestag's budgetary authority was challenged and whether the risks entailed by Germany as the largest guarantor were acceptable from a budgetary perspective. The sequence and pace of the Bundestag's involvement in crisis management decisions became a substantial issue between the government and the legislators. A number of parliamentarians took the matter to the Constitutional Court – with the result that the role of the German Bundestag was strengthened (see below). Parliamentarians were further concerned as to whether the packages were efficiently designed from an economic perspective. The necessary and preferable extent of financial solidarity in the common currency area was also widely debated, asking which adaptation efforts and governance reforms would be necessary in exchange.

Due to the nature of the Bundestag's concerns and its new assertiveness, there were moments when it had to be acknowledged that parliamentarians might veto a crisis decision. The government leveraged this potential veto player position along with the threat of a constitutional complaint to strengthen its bargaining position during European negotiations. However, Germany's parliamentary system, with efficient party structures and parliamentary groups, makes effective deadlock situations unlikely and the German government could rely on the support of pro-European opposition members and moral persuasion with regard to Germany's responsibility in Europe to get the majorities it needed. So, throughout the crisis, the government did get parliamentary approval of all crisis decisions.

Public Opinion and Political Communication

The concerns about developments in the euro area raised by parliamentarians and members of government reflected the considerable uncertainty as to whether German voters would punish political parties for their readiness to accept financial obligations that the rescue packages entailed. For instance, the government's reluctance to take earlier decisions on a rescue package for Greece in spring 2010 was partly related to the regional elections in North Rhine Westphalia, the most populous German state (Gatzke 2010). However, euro area issues were hardly discussed during regional election campaigns, one exception being the election in Berlin on 18 September 2011, when the increasingly unpopular FDP declared the federal state election to be a 'plebiscite on the euro' (Jungholt and Kensche 2011). Eurosceptic campaigning strategies did not improve electoral results in the regional elections and political parties articulating far-reaching demands for crisis management and a deepening of the euro area like the SPD and the Greens were not sanctioned by the voters for holding pro-European positions.

Two years into the sovereign debt and banking crises, German public opinion became critical of rescue packages, but did not turn anti-European. Even at the height of the crisis in 2011, 66% of Germans still supported 'a European economic and monetary union with one single currency, the euro' (European Commission 2011c). In a Eurobarometer poll, 88% thought that it was in the German interest to stabilise the euro area. In fact, greater integration was supported in various policy areas (European Commission 2011b). But the role of the euro during the crisis was not seen as particularly positive: Less than 40% of Germans thought that 'the euro has cushioned the effects of the economic crisis' (European Commission 2011a).

While public opinion remained comparatively stable, the German debate on the EU and the euro has become more polarised than it used to be. The more critical tones in the German debate until 2013 were not surprising, as the management of the debt crisis was not as successful as expected. Month after month, the government agreed to ever more far-reaching measures which it had excluded categorically only weeks before. The reform of the EFSF is one example. Also, the impression that the German government was taking on greater risks and that the rescue credits would not be paid back if recipient countries failed in

their reforms and consolidation measures led to growing scepticism.

In 2014, public opinion in Germany towards the EU and the euro has recovered and stabilised at a level above the EU average. According to poll data from the Pew Research Center, 66% of Germans are in 'favour of the EU' (six percentage points more than in 2013) and 72% of Germans support the euro. Besides, Germans have regained trust in European integration as an economic boost for Germany: a majority of 63% is convinced that Germany has benefited from European economic integration, 13 percentage points more than in the year of economic crisis, 2009, when only 50% shared this impression (Pew Research Center 2014). Public opinion in Germany can thus be characterised as broadly pro-European, matching the still prevailing pro-European consensus of the majority of German political elites. Nevertheless, the debate has become more polarised and the option of breaking up the monetary union, a taboo subject for more than a decade, has become part of it. The emergence of the AfD (*Alternative für Deutschland*), an outrightly Eurosceptic party that reached a vote share of 4.7% at the Bundestag elections in 2013 and 7% at the European elections in May 2014, will contribute to an increasing polarisation of the German debate on the European Union. German political leaders will increasingly have to justify Germany's political, economic and financial self-interest in further integration.

The Impact of the Constitutional Court

A major reason why the German government did not adopt a more proactive policy was the widely shared concern that rescue measures could be ruled unconstitutional by the Federal Constitutional Court. One argument was that a violation of the no bailout clause by one of the rescue packages would be problematic, as the German Constitutional Court had ruled that the Maastricht Treaty, which created the euro area, was in line with the German Constitution precisely because it contained that clause. Another argument, used by the complainants against the rescue package for Greece and the EFSF, was that their right to property and democratic legitimacy had been violated.

However, the Court so far has not stopped or modified any of the crisis management measures. In spring 2010, it refused the request of immediate action against the first rescue package for Greece because it did not see any evidence that the federal government's evaluation of the fiscal and currency situation had been deficient (Bundesverfassungsgericht 2010). On 7 September 2010, it also refused the constitutional complaint against the EFSF. The Court did not see a violation of the budgetary autonomy of the Bundestag and judged that the euro bailout fund is constitutional as long as the approval of parliament is given (Bundesverfassungsgericht 2011a). Furthermore, it did rule that the amount of the adopted grants did not exceed the limit of budgetary charge in such a way that budgetary autonomy had been compromised (Bundesverfassungsgericht 2011b).

The Court's decisions moreover led to a review of the participation of the Bundestag in European policymaking. For instance, the Court ruled the law on the 'committee of nine parliamentarians' unconstitutional, which, for reasons of efficiency, should be mandated to clear German approval for the use of the new crisis instruments of the EFSF. After the Court's decision on the Lisbon Treaty, this was the second time that the Federal Constitutional Court had requested that the Bundestag carry out its parliamentary control of European affairs more rigidly (Becker and Maurer 2009; Bundesverfassungsgericht 2012).

All in all, while the potential incompatibility of rescue measures with the German Basic Law was a strong limiting factor to the scope of potential steps envisioned by the government since the sovereign debt crisis hit Greece in 2010, the Court has never actually prevented the implementation of any European crisis management decision.

A Changing European Context Since 2012

The risk that Germany might not be able to ratify a European agreement – be it for political or legal reasons – gave the German government greater

leverage during negotiations in 2010 and 2011 on the pacing and design of crisis management measures, and on the shaping of the future governance structures of the euro area.

In 2012, the context changed. The notion of a hegemonic Germany which emerged in 2010/2011 was questioned, as the worst-case scenarios of a Constitutional Court ruling or a Bundestag decision blocking effective crisis management did not materialise. Moreover, German public opinion and the domestic debate on Europe did not turn overly anti-European or populist, and it meanwhile became increasingly clear that both German banks and industry had such a strong interest in a stable currency union that giving up the single currency was not an option. Although parts of Germany's exports shifted from the euro area to the world market (the euro area share of total German exports dropped from 43.8% in 2007 to 37.5% in 2012), the euro area and the EU remain crucially important export and import markets and FDI destinations for German companies. So, an exit threat or the scenario that Germany could decide not to support a crisis country lost credibility.

In addition, a number of European developments reduced Germany's relative political weight (Schwarzer and Lang 2012). First, the Franco-German relationship deteriorated after the election of François Hollande. In 2010/2011, Chancellor Merkel was able to predetermine European policy choices and discard other member states from the decision-making processes through her joint initiative with then-President Nicolas Sarkozy and her direct access to Council President Van Rompuy.

Since 2012, the North–South divide in the euro area has become a political dividing line. Shortly after his election, President Hollande tabled proposals for a new growth strategy, including project and euro bonds which the German government had previously excluded as an option. He sought closer contacts with the Italian and the Spanish prime ministers. The EU's southern member states have generally become more vocal, in particular as the growth performance of the programme countries provided arguments for calling Germany's austerity-first approach into question. After the European elections in 2014, the Italian Prime Minister Renzi, who was politically strengthened, has set out to challenge economic and budgetary policy making in the euro area.

Meanwhile, France became a much less active partner, with weaker political leadership and its own economic problems. The 2014 European elections, in which the right wing radicals of the Front National garnered a growing percentage of the vote, have further weakened the French political leadership.

Finally, since the ECB announced its potentially powerful bond-buying programme, OMT, Germany's role as crisis manager has become less crucial. At the time of heightened insecurity, related to the question of how crisis management and governance reform would evolve, Germany could impose policy choices on fellow member states that otherwise would not have secured the support of the majority, but which were accepted in order to overcome Germany's reluctance to engage more strongly in crisis management. With its 27% share in the rescue mechanisms, Germany's political commitment to the euro area is still systemically relevant. But when the ECB announced potentially unlimited bond purchases should the stability of the euro be endangered, it *de facto* reduced the importance of the government-led crisis management, and hence Germany's central role.

Future Perspectives

It is nevertheless likely that Germany will continue to push for further political integration, not least because doubts have arisen as to whether the strengthened rules-based coordination framework will actually yield the intended results. The German government is likely to support a further strengthening of fiscal and economic policy coordination. One political lesson of the recent rise of anti-EU and anti-establishment parties in the European elections is that stronger control of domestic policies should go hand in hand with stronger EU legitimacy. Germany has advocated an extension of competencies of the European Parliament for several decades. It remains to be seen whether it sticks to this position or whether it

argues for a strengthening of national parliaments. Germany is likely to seek inclusion of the ESM into EU Treaty law, possibly suggesting a further development of the ESM in the direction of a European Monetary Fund, including a stronger framework for dealing with sovereign defaults. Germany's fundamental position is likely to remain consistent: to minimise moral hazard as much as possible, to increase control and the credibility of rule-based coordination, to enable European interference with national policies, and to bring the market mechanism back in as a disciplining device.

For some time, Germany's influence in EU policymaking will continue to be ensured by its mere size and competitiveness. For the next three decades, Germany is likely to remain the largest country in the euro area with regard to population and economic weight. The next largest country, with 20% of GDP, is currently France. In the context of the EU 27, Germany's economy is more or less the same size as the combined economies of the 20 smallest member states. Moreover, its strong commitment to European integration and its current willingness and ability to shoulder larger financial burdens than other member states (through the EU budget, or more recently in the euro area rescue mechanisms) contributes to its relative strength in European negotiations. It will, however, be less able to determine the outcome of the process given the increasing polarisation in the EU. In this context, it will be crucial to reinvigorate Franco-German cooperation, especially if conceptual differences between both governments prevail. A revitalised bilateral cooperation has to be completed by involving various partners to buttress legitimacy and to create acceptance for leadership and the way to intensified integration. German governments will probably seek to include both euro area countries – to consolidate the common currency – as well as non-members, as one priority of Germany will continue to be the maintenance and strengthening of the single market. This is not only an important export market for Germany, but also the supplier market for successful German exporters. Competitiveness and the growth of fellow member states and the absence of currency fluctuations remain a German priority in order to back its drive for continued global competitiveness.

See Also

- ▶ European Central Bank
- ▶ European Central Bank and Monetary Policy in the Euro Area
- ▶ European Monetary Integration
- ▶ European Monetary Union
- ▶ European Union Budget
- ▶ Euro Zone Crisis 2010

Bibliography

Alexander, R. 2011. Kontrollverlust der Kanzlerin. Merkels historischer Krisen-Trip durch Berlin-Mitte. *Welt Online*, 22 October 2011. http://www.welt.de/politik/deutschland/article13675518/Merkels-historischer-Krisen-Trip-durch-Berlin-Mitte.html.

Becker, P., and A. Maurer. 2009. Deutsche Integrationsbremsen. Folgen und Gefahren des Karlsruher Urteils für Deutschland und die EU. *SWP-Aktuell* 41.

Bryant, C. 2012. Schäuble backs wage rises for Germans. *Financial Times*, 6 May 2012.

Bundesfinanzministerium. 2010. *Key points of the Federal Government for the strengthening of the euro zone*. http://www.bundesfinanzministerium.de/nn_53836/DE/Wirtschaft__und__Verwaltung/Europa/Der__Euro/20100520-Task-Force.html?__nnn = true.

Bundesverfassungsgericht. 2010. *Press release on the decision of 7 May 2010*. http://www.bverfg.de/pressemitteilungen/bvg10-030en.html.

Bundesverfassungsgericht. 2011a. *Decision of 7 September 2011*. http://www.bundesverfassungsgericht.de/entscheidungen/rs20110907_2bvr098710.html.

Bundesverfassungsgericht. 2011b. *Pressemitteilung Nr. 55/2011 vom 7*. September 2011. http://www.bundesverfassungsgericht.de/pressemitteilungen/bvg11-055.

Bundesverfassungsgericht. 2012. *Pressemitteilung Nr. 14/2012 vom 28*. Februar 2012. http://www.bundesverfassungsgericht.de/pressemitteilungen/bvg12-014.html.

Council of the European Union. 2010. *Press release. Extraordinary Council meeting on 9/10 May 2010*. https://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ecofin/114324.pdf.

Council of the European Union. 2011a. *Conclusions of the heads of state or government of the euro area*. https://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ec/119809.pdf.

Council of the European Union. 2011b. *Statements by the heads of state or government of the euro area and EU*

G

institutions. 21 July 2011. http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ec/123978.pdf.
Dullien, S. and Guérot, U. 2012. *The long shadow of ordoberalism. Germany's approach to the euro crisis*. ECFR Policy Brief 49, http://www.ecfr.eu/page/-/ECFR49_GERMANY_BRIEF_AW.pdf.
Dullien, S., and D. Schwarzer. 2011. *Dealing with debt crises in the Eurozone. Evaluation and limits of the European Stability Mechanism*. SWP research paper 11.
Dyson, K., and K. Featherstone. 1999. *The road to Maastricht. Negotiating economic and monetary union*. Oxford: Oxford University Press.
European Commission. 2011a. Annexe. *Standard Eurobarometer* 75.
European Commission. 2011b. Europeans, the European Union and the Crisis. *Standard Eurobarometer* 75.
European Commission. 2011c. Annexe. *Standard Eurobarometer* 76.
European Council. 2010. *Statement by the Heads of State or Government of the European Union at the informal meeting of the European Council on 11 February 2010*. http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ec/112856.pdf.
Gatzke, M. 2010. Merkel muss zur echten Europäerin werden. *Zeit Online*, 28 April 2010. http://www.zeit.de/wirtschaft/2010-04/merkel-griechenland.
Jungholt, T., and C. Kensche. 2011. Wahlkampf am Rande der Bedeutungslosigkeit. *Die Welt*, 16 September 2011 (own translation of quotation).
Pew Research Center. 2014. *A fragile rebound for EU image on eve of European Parliament elections*. http://www.pewglobal.org.
Schäuble, W. 2010. Why Europe's monetary union faces its biggest crisis. *Financial Times*, 11 March 2010.
Schwarzer, D., and K.-O. Lang. 2012. The myth of German hegemony. *Foreign Affairs* 2. http://www.foreignaffairs.com/articles/138162/daniela-schwarzer-and-kai-olaf-lang/the-myth-of-german-hegemony.

Germany, Economics in (20th Century)

Harald Hagemann

Abstract

The development of German economics in the 20th century is characterized by an interaction between internal scientific factors and external political factors. The dominance of the Historical School ended with the death of Schmoller and the First World War. The pressing economic problems of the young Weimar Republic stimulated macroeconomic research and national income accounting, whereas the Nazis' rise to power caused an important intellectual emigration from which German economics recovered only slowly after 1945. After an early period of ordoliberalism, as in other countries the development of economics has increasingly reflected a process of internationalization dominated by American economics.

Keywords

Arndt, H.W.; Aumann, R.; Bombach, G.; Brunner, K.; Bruno, M.; Burchardt, F.; Business cycles; Christaller, W.; Colm, G.; Corden, W.M.; Cowles Commission; Creative destruction; Economic liberalism; Eltis, W.; Entrepreneurship; Erhard, L.; Eucken, W.; Föhl, C.; Freiburg School; Game theory; General equilibrium; German Historical School; Germany, economics in; Giersch, H.; Great Depression; Hahn, F.; Harsanyi, J.C.; Hayek, F.A.; Hilferding, R.; Innovation; Jaffé, E.; Jessen, J.; Keynesianism; Krelle, W.; Launhardt, C.F.W.; Lautenbach, W.; Lederer, E.; Leontief, W.; Liefmann, R.; Lösch, A.; Lowe, A.; Mandelbaum, K.; Marginal Utility School; Marschak, J.; Marx, K. H.; Meidner, R.; Menger, C.; Menger, K.; *Methodenstreit*; Mises, L. von; Monetarism; Morgenstern, O.; Musgrave, R.; Myrdal, G.; Nash, J.; National income accounting; Neisser, H.; Neoclassical synthesis; Ordoliberalism; Peter, H.; Preiser, E.; Price control; Protectionism; Röpke, W.; Roscher, W.; Rosenstein-Rodan, P.; Schelling, T.; Schmoller, G. von; Schneider, E.; Schumpeter, J.A.; Selten, R.; Singer, H.; Social market economy; Sombart, W.; Spatial economics; Spiethoff, A.A.K.; Stackelberg asymmetric duopoly; Stackelberg, H. von; Stagflation; Streeten, P.; Tisch, C.; Wage controls; Wagemann, E.; Weber, A.; Weber, M.; *Werturteilsstreit*

JEL Classifications

B2

1900–1918: Dominance of the Historical School

Until the beginning of the 20th century there were relatively few economics students in Germany and their training heavily depended on other disciplines. Not until the 1880s was a special Ph.D. option opened for economists, and it was not until 1923 that a special curriculum was set up that led to a diploma degree. Nevertheless, in the second half of the 19th century a growing professionalization process took place, with the emergence of scholarly journals and the foundation of economic societies, the most important of which was the Verein für Sozialpolitik in 1872 (Hagemann 2001). The driving force behind, and chairman of, the Verein from 1890 until his death was Gustav Schmoller (1838–1917), the undisputed leader of the 'Younger' German Historical School and the most influential economist in imperial Germany, particularly in Prussia, at the beginning of the 20th century. Schmoller favoured a historical and ethical approach to economics and the inductive method of collecting large amounts of statistical material, rather than the abstract axiomatic–deductive method of the classical economists and his neoclassical contemporaries. From 1883 he was involved in a famous dispute on method with Carl Menger; however, although a major issue in German literature, this *Methodenstreit* did not play a significant role at the meetings of the Verein. In remarkable contrast, the question of whether economists or other social scientists should make normative judgements had been a heated issue since the 1880s. This *Werturteilsstreit,* in which Max Weber, with his strong plea for a clear separation of social *science* from social *policy*, was a key figure, escalated at the 1909 Vienna meeting. 'The controversy about norms and values' raged until the outbreak of the First World War, at which time it was still unsettled.

1918–33: The Weimar Republic and the New Complexity

After the end of the First World War, and with the death of Schmoller in 1917, the Historical School lost the dominant position it had acquired in the German Empire from 1871, although it remained influential among many economists. In the early years of the Weimar Republic urgent policy problems, such as the socialization of production, reparation payments and particularly hyperinflation, dominated. From the mid-1920s onwards, theoretical and empirical research on business cycles became the most important issue, and a new generation of more theoretically oriented young economists came to the fore. This is best reflected in the 1928 Zurich meeting of the Verein für Sozialpolitik which focused on the explanation of business cycles. Major contributions were delivered by brilliant young economists including Adolph Lowe (born 1893), Friedrich August Hayek (born 1899), Wilhelm Röpke (born 1899) and Oskar Morgenstern (born 1902).

The long-run development of the capitalist economy and the study of the crises and economic fluctuations surrounding it had been a major issue since the mid-19th century. This holds for representatives of the 'Older' Historical School such as Wilhelm Roscher as well as for Karl Marx. At the beginning of the 20th century leading economists like Arthur Spiethoff and Werner Sombart became widely known with their studies on business cycles, with Joseph A. Schumpeter as the towering figure. In his *Theory of Economic Development* (Schumpeter 1911/12) the idea of creative destruction by innovation and the notion that bank credit is the prerequisite of innovation and of the foundation of new enterprises are of central importance. According to Schumpeter, capitalist development can proceed only in a wavelike form, and pioneering entrepreneurs are the key agents for this development.

More systematic empirical research on business cycles became established with the foundation of the German Institute for Business Cycle Research in Berlin in 1925, with Ernst Wagemann as director (he was also President of the Statistisches Reichsamt, the German Statistical Office). Two years later the Austrian Institute for Research on Business Cycles was founded in Vienna on the initiative of Ludwig von Mises. Hayek, the first director, was succeeded by Morgenstern in 1931 after Hayek moved to the

London School of Economics. Furthermore, in 1926 the Kiel Institute of World Economics engaged Adolph Lowe to build up a new department for statistical world economics and international trade cycles, which soon developed a most promising research programme. Lowe managed to recruit a group of highly talented young economists including Gerhard Colm, a leading expert on public finance and financial statistics (which was especially important for the payment of reparations), who later became the chief architect of West Germany's successful currency reform of 20 June 1948; the monetary theorist Hans Neisser; Fritz (later Frank) Burchardt, who after the Second World War became director of the Oxford Institute of Statistics; Wassily Leontief (1927–28, 1930–31) who, while at Kiel, wrote his Berlin Ph.D. thesis on the economy as a circular flow but mainly worked on the statistical analysis of supply and demand curves; and Jacob Marschak (1928–30). Lowe's Kiel habilitation thesis, 'How is business cycle theory possible at all?' (Lowe 1926), which raised the fundamental methodological problem of the (in) compatibility of business cycle theory with the theory of general economic equilibrium, also exerted a major influence on Hayek, as can be seen from the latter's 1928 Vienna habilitation thesis, published in English as *Monetary Theory and the Trade Cycle* (1933, ch. 1).

The pressing economic problems of the early Weimar Republic and the mass unemployment and deflation of the Great Depression in the closing years of the republic, which required practical solutions, can also explain why German economics in the interwar period was not completely dominated by academic economists. Practitioners at public and private research institutions as well as leading bureaucrats played a key role. This is reflected, for example, in the core members of the Kiel group, Lowe, Colm and Neisser, who all had worked for the Weimar government and/or the Statistisches Reichsamt, or in the personage of Wilhelm Lautenbach, a leading member of the Ministry of Economics whose practical proposals during the Great Depression earned him the nickname 'the German Keynes'. Due to governmental needs the Weimar Republic invested heavily in macroeconomic research, which brought it to the forefront of statistical innovation and the development of national income accounting. Wagemann, as the key figure, provided the subsequent Nazi government with the statistical tools for economic planning (Tooze 2001).

1933 and After: Dismissal, Expulsion, and Emigration of Germanspeaking Economists

The political events of 1933 marked an important turning point for the economics profession in Germany. Shortly after its rise to power the new Nazi government passed the Restoration of Civil Service Act, which formed the basis for the dismissal of 'disagreeable' persons from public services either for racist or for political reasons. By the winter of 1934/35 about 14 per cent of the faculty at German universities had been dismissed, but in economics the figure was 24 per cent. However, the dispersion was significant. Whereas some years later in Austria the great majority of dismissals and expulsions were concentrated on the University of Vienna, in Nazi Germany the three universities of Berlin, Breslau (today's Wroclaw in Poland) and Frankfurt had to cope with the greatest losses. Breslau in Silesia had a large and recognized Jewish community – it was home, for example, to Fritz Haber, the Nobel Prize-winner in chemistry, and the historian Fritz Stern (1999). Neisser, the development economist Heinz Wolfgang Arndt, the trade theorist Warner Max Corden and the later architect of the Swedish workers' investment funds Rudolf Meidner, all came from Breslau.

In economics, the highest losses were suffered by the faculties of Heidelberg (which lost seven of 11 members), Kiel (five of ten) and Frankfurt (13 of 33). Heidelberg had been a blossoming centre of liberal intellectual discourse in the Weimar Republic and characterized by multidisciplinarity, as expressed in Emil Lederer's analysis of the 'new middle classes'. A key role there was played by the Institute for Social and State Sciences, founded in 1924, with Max Weber's younger brother Alfred as the director

until 1933 when he ordered students to remove the swastika flag from the main campus building. (Alfred Weber can be regarded as one of the very few scholars for whom the notion of 'internal emigration' really fits.) The new Goethe University in Frankfurt had developed into a leading centre of the social sciences within the short period of the Weimar Republic, as can be seen by looking at the long list of outstanding scholars dismissed in 1933, including Franz Oppenheimer, Karl Pribram, Lowe, Fritz Neumark, Karl Mannheim and his research associate Norbert Elias, the theologian Paul Tillich, among others. When Gunnar and Alva Myrdal visited the Institute of World Economics in Kiel in the summer of 1933 on behalf of the Rockefeller Foundation they diagnosed a deteriorated scientific reputation because by then almost all the best scholars had emigrated.

The long-term loss of quality and of international reputation of German economics caused by the political events of 1933 and thereafter is also reflected in the evolution of scholarly journals. German-language journals not only lost most of the émigré economists as contributors, but also most foreign economists stopped writing in the German language or publishing in German (and from 1938 onwards also in Austrian) journals. The obverse of this was the increased number of articles written by émigré economists in the leading Anglo-Saxon journals. With the exception of Spiethoff, who stayed in office as the editor of *Schmollers Jahrbuch,* journal editors were replaced after the Nazis' rise to power (Hagemann 1991). From 1904 the *Archiv für Sozialwissenschaft und Sozialpolitik* had been the leading journal in economics and the social sciences, with a run of fine editors, Max Weber, Werner Sombart and Edgar Jaffé, followed in 1922 by Emil Lederer and his two associates, Joseph Schumpeter and Alfred Weber. But this journal too had to cease publication under Nazi rule. For a short period in the 1930s the Vienna-based *Zeitschrift für Nationalökonomie* became the outstanding scholarly journal in economics in the German language, under the experienced editorship of Oskar Morgenstern. It published important articles particularly on capital theory, the role of time in economics or general equilibrium analysis, but after Hitler's invasion of Austria the quality of the journal collapsed. In the wake of the *Anschluss* the Vienna Institute for Research on Business Cycles became a branch office of the Berlin Institute.

The group of dislocated German and Austrian economists who had acquired academic degrees comprises 253 scholars, of whom 148 were dismissed from universities, 57 from private research institutes, and 28 from other public employment, and 20 were young economists who just had completed their studies, students like Richard Musgrave who had gained his diploma degree at the University of Heidelberg in May 1933 (Hagemann and Krohn 1999). Some 221 (87 per cent) emigrated. Of the remaining 32 several were killed in the Holocaust, concentration camps, or Gestapo prisons, these including Rudolf Hilferding, Robert Liefmann and Cläre Tisch. The intellectual loss included 75 members of the so-called 'second generation', young students or pupils who emigrated with their parents and later made a career as economists, such as, for example, Robert Aumann, Michael Bruno, Otto Eckstein, Walter Eltis and Frank Hahn.

1933–45: German Economics in the Nazi Period

German economics in the Nazi period moved far away from economic liberalism, and state interventions and regulations such as price controls and wage freezes played a major role (Janssen 2000). The 1932 Dresden meeting of the Verein für Sozialpolitik, organized by Mises, focused on problems of value theory and was dominated by the representatives of the Marginal Utility School, but was also clearly marked by heated debates on protection and the question of autarky. This meeting turned out to be the last before the outbreak of the Second World War because in December 1936 the majority of the members decided to dissolve the Verein in order to escape *Gleichschaltung* by the Nazis.

This decision shows that the group of dedicated National Socialists, which included Feder

and Wiskemann, was remarkably small. More important in numbers and influence were the two groups of fellow travellers including Gottl-Ottlilienfeld, Predöhl, Wagemann and national and conservative opportunists including Sombart and Spann respectively (Rieter and Schmolz 1993, p. 95), who largely followed the historical–holistic approach and also introduced ideas from contemporary German philosophy into economics. A fourth group consisted of 'renegades', former Nazis who later changed their views. Prominent members of this latter group include Jens Jessen – who in 1933 succeeded Harms as the director of the Kiel Institute and, as a result of his involvement in the failed attempt to assassinate Hitler in July 1944, was executed some months later – and Heinrich von Stackelberg, who probably only escaped the same fate because of his move to Spain in 1943. A fifth group consisted of opponents of the Nazi regime who either passively distanced themselves from the regime, thereby ending their professional careers, or actively fought against it, like the members of the Freiburg School.

The first subgroup of opponents included Hans Peter, a very able mathematical economist and theorist of the circular flow, who together with Erich Schneider and Stackelberg from 1935 to 1942 edited the new journal *Archiv für mathematische Wirtschafts- und Sozialforschung.* Due to his defence of his liberal socialist convictions in the Nazi period Peter obtained a full professorship at the University of Tübingen only in 1947. August Lösch, a brilliant economist of great personal integrity, received his habilitation from the University of Bonn in 1939 with his *The Economics of Location*, in which he applied general equilibrium theory to the space dimension. (Since the days of Thünen, then via Launhardt, Alfred Weber and Christaller to Lösch, spatial economics has been an area of economics where German economists have made important contributions.) Lösch himself, however, did not have a successful professorial career: because of his outspoken anti-Nazi views he survived fascism only by becoming a researcher at the Kiel Institute. He died tragically from scarlet fever a few weeks after the end of the war.

Although in international terms German economics fell behind in the Nazi period, there were nevertheless some significant contributions. Stackelberg was one of the most gifted theoretical economists. His *Marktform und Gleichgewicht* (1934), with its creation of the Stackelberg asymmetric duopoly, went beyond the work of Chamberlin and Joan Robinson in the depth of its theoretical analysis and in its mathematical rigour. However, although he was one of the very few German economists whose analysis was deeply embedded in the Anglo-Saxon approach to price and cost theory, and although the book was immediately reviewed by top theorists such as Hicks, Kaldor, Lange, Leontief and Zeuthen, it failed to make an impact. Because its author was a well-known supporter of the Nazis and the book ended with a short paean to the corporate state, the book's theoretical achievements were overlooked and even today it has not been translated into English.

Keynes was a central figure of reference in theoretical and policy debates in interwar Germany ever since his opposition to the reparation stipulations of the Versailles Treaty; not surprisingly, the first translation of his *General Theory* was into German (with a special Preface published in the same year 1936) and was extensively reviewed and debated. Leaving aside the complex question of German anticipations of the *General Theory* (Bombach et al. 1981), Carl Föhl's *Geldschöpfung und Wirtschaftskreislauf,* published in 1937 but already completed in December 1935, exhibits striking parallels with Keynes's theories (despite its using a different conceptual apparatus), and was the outstanding achievement in contemporary German literature on macroeconomics.

The Post-1945 Development of Economics in Germany: Ordoliberalism and the Social Market Economy

As is well known, after the Second World War, economic order and economic policy in the new Federal Republic of Germany were decisively

influenced by the ordoliberal thinking of Walter Eucken and the Freiburg School and the principles of the social market economy (Watrin 1979). However, the roots of ordoliberalism go back to the years 1938–45, with opposition to National Socialism based on Christian convictions (Rieter and Schmolz 1993). Eucken's main work, *Foundations of Economics,* was published in 1940. Although he was a well-known critic of the Historical School, his taxonomic approach, which focuses on reconciling economic with legal, institutional and social factors, is clearly embedded in the German tradition of state sciences. Thus in the competitive order he perceived that the state plays a substantially stronger role than his Anglo-American colleagues would allow, not to speak of Austrian contemporaries such as Mises and Hayek. This is particularly visible in Eucken's 'regulating principles' – that is, monopoly control, social policy, external effects – that supplement the 'constituting principles' of the market economy, that is, private ownership, competition in open markets, freedom of contract, a functioning price mechanism, monetary stability and consistency in economic policy.

German economics, which went through a laborious catching-up process after 1945 without being able to compensate fully in the following decades for the loss of qualified personnel in the Nazi period, suffered a further blow by the untimely deaths of Lösch (1945), Stackelberg (1946) and Eucken (1950).

The economists who were driven out of Nazi Germany had meanwhile contributed significantly to innovative research in their host countries. This holds particularly for the United States, which was the final destination for about 60 per cent of the émigré economists, but also for the UK where the new field of development economics had been shaped by Paul Rosenstein-Rodan, Kurt Mandelbaum (Martin), Arndt, Hans Singer, Paul Streeten and others, all émigrés from German-speaking countries. In the United States economists at the New School for Social Research, where the graduate faculty had been founded as the 'University in Exile' in autumn 1933, gained a certain influence in the period of the Roosevelt and Truman presidencies (Krohn 1993), but in the long run those who kept their distinctly German scholarly identity were less influential.

Among the important new developments in economics which were transferred to Germany after 1945, if with some delay, were game theory and econometrics, developed in particularly at the Cowles Commission in Chicago after Marschak became the Research Director there in 1943. This work was later mirrored in Wilhelm Krelle's research centre in econometrics at the University of Bonn in the 1960s and 1970s. The foundations of modern game theory were laid by John von Neumann and Oskar Morgenstern's *Theory of Games and Economic Behavior* (1944), which was the fruit of their cooperation at the Institute for Advanced Study in Princeton in the years 1939–43. Their work together came about as a result of exile, although the Budapest-born mathematician von Neumann had made his habilitation at the University of Berlin in 1927 and presented his famous paper on the general economic equilibrium of an expanding economy at Karl Menger's mathematical colloquium at the University of Vienna in 1936. The award of the Nobel Prize in economics in 1994 to John F. Nash, the Hungarian-born John C. Harsanyi, and Reinhard Selten as the first German economist, as well as in 2005 to Robert Aumann and Thomas Schelling, also reflects the great role of German, Austrian and Hungarian scholars in the development of game theory.

The development of modern public finance by Musgrave reflects the crossfertilization of the more theoretically oriented and rigorous Anglo-Saxon tradition of public finance with the German tradition of *Finanzwissenschaft*, including its institutional, historical and legal aspects. With his division of the public sector into three branches, which besides allocation also include stabilization and distribution as a fiscal concern, Musgrave indicates some German influences in the émigré's baggage (Musgrave 1996). The German translation of Musgrave's *Theory of Public Finance* (1959) was used at most universities for more than two decades as the standard textbook. Distributional theory and wealth formation among workers were also key issues for some

leading German economists – Erich Preiser, Gottfried Bombach and Krelle – in the 1960s.

1967–74: The High Years of Keynesianism

At German universities the Keynesianism of the Hicks–Samuelson neoclassical synthesis had become the dominant position since the late 1950s. This can largely be attributed to Erich Schneider, whose three-volume *Introduction into Economic Theory,* originally published in 1946–52, became the dominant textbook in the 1950s and 1960s, going through many editions. Schneider, who had made his habilitation with Schumpeter in Bonn in 1932 and became Professor in Aarhus in 1936, came back from Denmark in 1946 to become professor in Kiel, where he also directed the Institute of World Economics from 1961–9 (when he was succeeded by Herbert Giersch). From 1963 to 1966 Schneider was chairman of the Verein für Socialpolitik, which had been re-founded in 1948. When the influential Theoretical Committee was re-established shortly afterwards, Schneider exercised his power as chairman in the direction of a more mathematically oriented approach, which at the beginning had to overcome strong resistance (Schefold 2004).

Until the recession of 1966–7, however, economic policy was still dominated by ordoliberal ideas. As a consequence, Ludwig Erhard, who had been a successful Minister of Economics from 1949 to 1963, lost his job as Chancellor in December 1966, when the first 'Grand Coalition' of Christian and Social Democrats was formed. With the Social Democrats' entry into government and the ratification of the Stability and Growth Act in June 1967, Keynesianism gained a relatively late admission into Germany. According to Article 1 of the Act, the federal and state governments

> have to respect the requirement of macroeconomic equilibrium in their economic and financial policy measures which have to be taken in a way that they contribute, within the scope of a market economy, to simultaneously achieve stability of the price level, a high level of employment, and external equilibrium together with steady and appropriate growth.

These four macroeconomic goals appeared in the statutes of the German Council of Economic Advisers (CEA), which was founded in August 1963 and from autumn 1964 presented its annual report. The German council differs from the American in being an external and independent committee for policy consultation rather than part of the government.

In the public eye Karl Schiller's term of office as economics minister from the end of 1966 to the summer 1972 is remembered as the heyday of Keynesian economic policy in Germany. This is due to Schiller's remarkable ability to coin phrases such as *'Globalsteuerung'* ('macroeconomic demand management'), and his charismatic interpretation of economic policy which contributed to a widespread belief in the government's management power of macro variables, before the first oil price shock and the new phenomenon of stagflation shook that confidence. However, it should not be overlooked that Schiller had always followed a synthesis of Keynesianism and ordoliberal ideas. This is expressed most clearly in his influential article on economic policy in the *Handwörterbuch der Sozialwissenschaften* (Handbook of Social Sciences) (Schiller 1962), in which he formulated his famous credo: 'competition to the extent possible, planning to the extent necessary', with 'planning' understood in the sense of Keynesian demand management. Through his homage to Eucken, that is, in supplementing process policy with *Ordnungspolitik,* Keynesian policies took on a distinctly German tinge. This came against the background of discrediting the interventionist policies of the Nazi period, policies that were being pursued in Stalinist East Germany, and the need to safeguard the market economy against Marxist policies that were finally given up by the Social Democratic Party (SPD) only in its Godesberg programme adopted in 1959. In line with Giersch and the majority of the CEA, Schiller also advocated flexible exchange rates in the final years of the Bretton Woods system and in debates in the German 1969 election campaign when currency flexibility was heavily opposed by the Christian Democrats and the German export industry. The strong revaluation of the

Deutschmark thereafter contributed to a dampening of inflation in Germany.

From 1974 to the Present: Is There a German Economics?

Even in the short period when Keynesian influence on policy was at its peak, the Bundesbank was a powerful institution that followed its own policy of securing price stability, thereby constraining the implementation of Keynesian full employment policies. After the second oil price shock, when the German economy ran into a current account deficit in 1979–81 and there was unusual pressure to devalue the Deutschmark, the Bundesbank reacted by raising interest rates, a restrictive monetary policy that led to a major controversy with Chancellor Schmidt. After December 1974, when for the first time it had announced a target for the growth of the money supply, the Bundesbank followed an explicit monetarist policy, in line with many of the major central banks in the Western world. The Constance seminar on monetary theory and policy, initiated by Karl Brunner, which from 1970 had brought together American and German economists as well as practitioners from the Bundesbank and commercial banks, was the main vehicle for the breakthrough of monetarist ideas in theory and practice. Since 1976 the CEA, the majority of whose members had favoured moderate wage policies – which repeatedly led them into controversy with the trade unions – explicitly propagated supply side policies.

With German unification in 1990 another German *Sonderweg* ended. University economics faculties in East Germany, where from 1945 to 1989 economics had been mainly reduced to Marxism-Leninism and a narrowly defined socialist business administration, were now completely restructured, as also happened in law and the social sciences. The government tried to restore the pre-1933 prestige of the Humboldt University in Berlin, formerly Germany's leading academic institution, which produced many Nobel Prize-winners in physics, chemistry and medicine. Its economics faculty entered the club of the leading faculties, which included Bonn, Mannheim and Munich.

The post-1945 development of economics is characterized by a growing process of internationalization combined with an increasing dominance of American economics (Coats 1997, 2000). The triumphant ascent of American economics after the Second World War is the consequence of the economic and political leadership of the United States, the benefits of the importation of scholars from Hitlerian and Stalinist Europe – Scherer (2000, p. 622) calculates the citations received by the German-speaking émigré economists in the Social Science Citation Index (SSCI) for the period 1960–4 as 'roughly equivalent to the adjusted citation output of the first- ranked Harvard and second-ranked MIT plus the 19th-ranked University of Illinois economics departments' – and a national style of economic research characterized by the early introduction of graduate studies at the leading universities, with pressure to acquire the necessary mathematical and econometric tools for a specialized theoretical and applied work. The consequential international convergence process and the 'professional *Gleichschaltung'* (Peacock in Frey and Frey 1995, pp. 267–71) associated with it led to increasing debates of the type 'Is there a European economics?' (Frey and Frey 1995). In that sense, at the end of the 20th century there is no recognizable 'German' economics, as there was at the beginning of the century. German economists participate in international networks as do economists from other European countries or other parts of the globe, with English as the lingua franca. This is also reflected in the fact that since 2000 the Verein für Socialpolitik has published the international *German Economic Review* instead of the *Zeitschrift für Wirtschafts- und Sozialwissenschaften,* formerly *Schmollers Jahrbuch,* which had been revived under its old name but is no longer linked to the Verein, which in 2000 had 2,928 individual members from more than 20 countries.

Bibliography

Bombach, G., K.-B. Netzband, H.-J. Ramser, and M. Timmermann, eds. 1981. *Der Keynesianismus III. Die geld- und beschäftigungstheoretische Diskussion in Deutschland zur Zeit von Keynes*. Berlin/Heidelberg/New York: Springer.

Coats, A.W. 1997. *The post-1945 internationalization of economics*. Durham: Duke University Press.
Coats, A.W. 2000. *The development of economics in Western Europe since 1945*. London: Routledge.
Eucken, W. 1940. *The foundations of economics*, 1950. London: William Hodge.
Föhl, C. 1937. *Geldschöpfung und Wirtschaftskreislauf*. Berlin: Duncker & Humblot.
Frey, B.S., and R.L. Frey. 1995. Is there a European economics? *Kyklos* 48: 185–311.
Hagemann, H. 1991. Learned journals and the professionalization of economics: The German language area. *Studi Economici* 1: 33–57.
Hagemann, H. 2001. The Verein für Sozialpolitik from its foundation (1872) until World War I. In *The spread of political economy and the professionalisation of economists: Economic societies in Europe, America and Japan in the nineteenth century*, ed. M.M. Augello and M.E.L. Guidi. London/New York: Routledge.
Hagemann, H., and C.-D. Krohn. 1999. *Biographisches Handbuch der deutschsprachigen wirtschaftswissenschaftlichen Emigration nach 1933*, 2 vols. Marburg: Metropolis.
Hayek, F.A. 1929. *Monetary theory and the trade cycle*. London: J. Cape, 1933.
Janssen, H. 2000. *Nationalökonomie und Nationalsozialismus. Die deutsche Volkswirtschaftslehre in den dreißiger Jahren*. Marburg: Metropolis.
Krohn, C.-D. 1993. *Intellectuals in exile: Refugee scholars and the new school for social research*. Amherst: University of Massachusetts Press.
Lösch, A. 1939. *The economics of location*, 1954. New Haven: Yale University Press.
Lowe, A. 1926. How is business cycle theory possible at all? *Structural Change and Economic Dynamics* 8(1997): 245–270.
Musgrave, R.A. 1996. Public finance and Finanzwissenschaft traditions compared. *Finanzarchiv* 53: 145–193.
Rieter, H., and M. Schmolz. 1993. The ideas of German Ordoliberalism 1938–45: Pointing the way to a new economic order. *European Journal of the History of Economic Thought* 1: 87–114.
Schefold, B. 2004. Wissenschaft als Gegengabe – Neugründung und Aktivitäten des Theoretischen Ausschusses im Verein für Socialpolitik von 1949–1973. *Schmollers Jahrbuch* 124: 579–608.
Scherer, F.M. 2000. The emigration of German-speaking economists after 1933. *Journal of Economic Literature* 38: 614–626.
Schiller, K. 1962. *Wirtschaftspolitik. Handwörterbuch der Sozialwissenschaften*, vol. 12. Stuttgart: Gustav Fischer.
Schneider, E. 1946. *Einführung in die Wirtschaftstheorie*, 3 vols. Tübingen: Mohr-Siebeck.
Schumpeter, J.A. 1911. *The theory of economic development: An inquiry into profits, capital, credit, interest, and the business cycle*. Cambridge: Harvard University Press, 1934.
Stackelberg, H. von. 1934. *Marktform und Gleichgewicht*. Vienna/Berlin: Julius Springer.
Stern, F. 1999. *Einstein's German World*. Princeton: Princeton University Press.
Tooze, A. 2001. *Statistics and the German state, 1900–1945*. Cambridge: Cambridge University Press.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.
Watrin, C. 1979. The principles of the Social Market Economy: Its origins and early history. *Zeitschrift für die gesamte Staatswissenschaft* 135: 405–425.

Germany's Historical Relationship with the European Union

Daniela Schwarzer

Abstract

This article gives an overview of Germany's European policies since the onset of European integration in the early 1950s. It covers German perspectives and Germany's impact on a variety of European policy areas with a special focus on the European economic order. After the Second World War, Germany was politically and economically isolated and under the external control of the Western Allies. In this historical context European integration and embeddedness helped Germany to rehabilitate itself as a political actor and to boost trade and economic growth. Inside the European framework and in close cooperation and coordination with France, Germany has exerted considerable leadership in shaping the European order and European institutions. Germany has managed to include national models, norms and policy principles in crucial European projects like the single market, including competition policy, monetary and economic union and eastern enlargement. In the course of the 1990s Germany underwent a process of 'normalising' its EU and foreign policy, transforming itself into a more pragmatic and self-confident power that

no longer hesitates to act explicitly in its national interest, outside a multilateral setting if necessary. This process culminated in Germany's controversial leadership role in the sovereign debt crisis in 2010.

Keywords

Economic governance; Economic order; European Union; EU governance; Euro; Euro area; European integration; France; German history; Germany; Leadership; Single market

JEL Classifications

E6; E42; F15; F33; F36; H3; H6

Germany's Historical Relationship with the European Union

Since the 1950s, all German governments have been committed to European integration. Yet Germany's perspectives on the European Union and its own role in the EU have evolved considerably over that period. In the immediate post-war years, the defeated and non-sovereign Federal Republic of Germany identified its integration into European structures as the only viable way to become a stable and prosperous democracy. After regaining strength and domestic political stability, Germany became an engine of European integration, alongside France, and considerably shaped the EU as we know it today. Since the start of the sovereign debt crisis in 2010, Germany's role has given rise to a controversial debate. For example, it has been characterised as a reluctant leader, a quasi-hegemon or even a colonial power imposing policy choices on other member states, while some critics have cast doubt on Germany's European commitment altogether. This article provides an overview of Germany's perspectives on European integration and its role in the process over six decades, focusing in particular on its impact on shaping the economic order of the European Union.

Germany After the End of the Second World War

After the Second World War, the defeated and divided Germany was politically isolated and economically devastated. Its industrial output was reduced to a third of its pre-war level, while its workforce had shrunk due to extensive losses of around 1.8 million people (or 3.3% of the total population) during the war. Food production was half of its pre-war level, resulting in starvation in the population (Henderson 2007; Wehler 2009). Moreover, 20% of housing was destroyed; even more in larger cities.

According to the Occupation Statute of 1949, Western Germany was not a sovereign state. Its external representation was carried out by the Allied High Commissioners of France, the UK and the USA. The Federal Republic was subject to economic discrimination and far-reaching restrictions. The Occupation authorities shared the responsibilities of representing German economic interests abroad, ratifying international treaties as well as managing external trade and foreign exchange. The Western Allied forces also controlled two of Western Germany's main industrial areas, the Ruhr and Saar areas, leaving it little leeway in choosing its own economic policy.

The Federal Republic was also denied having either its own army or its own arms industry, at a time when it faced growing threats to its security and territorial integrity as a consequence of increasing tensions among the Western Allied Powers of France, the USA and the UK on the one hand, and the Soviet Union on the other. The Western German political leadership was highly aware of its dangerous position between two ideologically driven nuclear superpowers and its dependence on the presence of Western Allied troops on German territory to provide security, for which it was contractually obliged to pay occupation costs of around DM7 bn per year (Dernburg 1955, p. 651).

Rehabilitating and Protecting Germany Through Integration into the West

Against this background, the prime objective of Germany's post-war political leadership was to provide some perspective of economic recovery to the country while assuaging the West's fears of a possible re-strengthening of post-Nazi Germany and maintaining the security guarantees provided by the Western Allies. The first Chancellor of the Federal Republic of Germany, the Christian Democrat Konrad Adenauer, perceived European integration as the best way to overcome Germany's isolation within Europe. When the French government proposed the creation of the European Coal and Steel Community (ECSC) in 1951, Adenauer reacted very favourably, in particular because he considered establishing a close and resilient relationship with France as a prerequisite for integration with the West (*Westbindung*).

The creation of the ECSC laid the foundation for further steps of sectoral integration, each of which the government of the Federal Republic supported. Only six years later, the Treaties of Rome established the European Economic Community (EEC) and Euratom. German membership in the European Communities, together with the other five founding members (France, Italy and the Benelux countries), was supported by a broad political consensus across German political parties, initiating a "virtuous circle that transformed the Federal Republic into a stable, liberal-democratic state embedded at the heart of a wider (West) European stability" (Jeffery and Paterson 2003, p. 61).

In parallel with the first steps of economic integration in Europe, Germany also became part of the transatlantic security structures. Constraints on the West German military ended on 5 May 1955, when the Paris Treaties became effective and lifted the occupation statute. The Federal Republic joined NATO and the Western European Union and became, in return, a sovereign state, entitled to its own foreign policy and foreign trade. (The division between West Germany (the Federal Republic of Germany) and East Germany (the German Democratic Republic), however, persisted until German reunification. Full German sovereignty was only restored with the Two Plus Four Treaty in 1990.) The treaties fixed the maximum size of the new German army (*Bundeswehr*) at 500,000 men, and obliged Germany to contribute naval and air forces for common defence purposes as well as placing 12 military divisions under NATO command (Dernburg 1955, p. 648; Federal Ministry of Defence 2013). Furthermore, Germany agreed to refrain from developing nuclear and chemical weapons. As a result of post-war demilitarisation, its strong dependence on external security guarantees and its historical war experience, Germany perceived itself as a civilian power, a notion which still prevails today with the German political elite and public.

Enabling Germany's Post-War Economic Miracle

As in the field of foreign and defence policy, the 1950s proved to be a turning point for Germany in economic terms. After the end of the war, the German economy suffered from repressed inflation. The Reichsmark was often not accepted in transactions, which hampered the efficiency of the economic system. Due to the low value of the currency as well as constant shortages of food and consumer goods, bartering was widespread in business transactions.

The currency reform of 1948, however, brought down inflation and price controls ended. At the same time, the Marshall Plan was put into place, combining two instruments: the European Recovery Programme (ERP) and economic reconstruction on the basis of sovereign debt relief and trade integration. Direct financial aid through ERP was comparatively less important for the German economy than the act of debt forgiveness. Between 1948 and 1952, the Federal Republic received US$1.4 billion, amounting to between 2% and 4% of West German gross domestic product (GDP) at the time (Ritschl 2012; Wehler 2009, p. 55). Creditor claims on Germany were frozen, as recipients of Marshall Aid were shielded from all foreign claims as long as these loans had not been paid back. The restructuring of debt meanwhile brought down Germany's public debt from

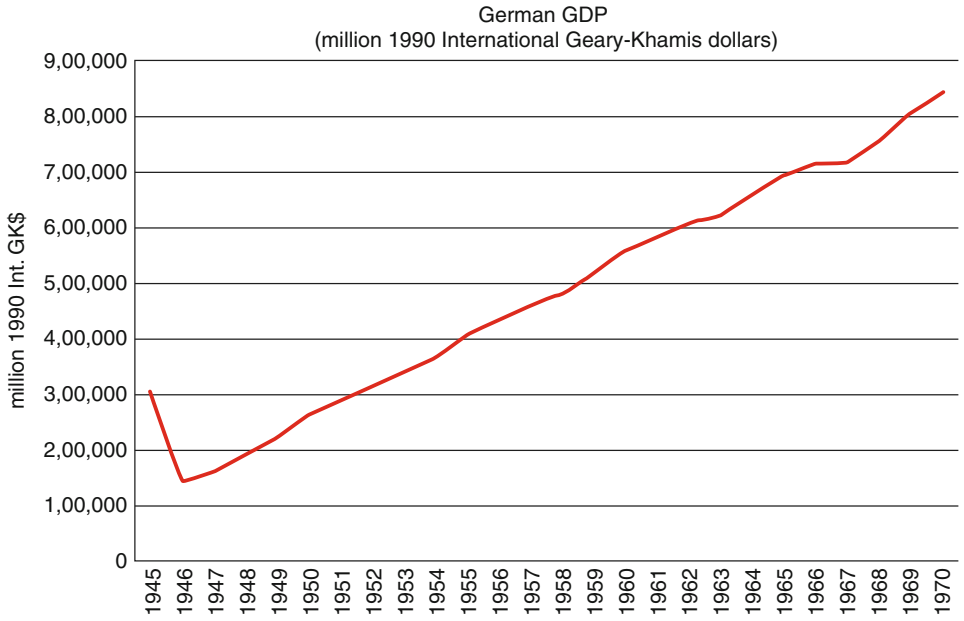
300% of its GDP to less than 20%, while the rest of Western Europe struggled with ratios near 200% throughout the 1950s.

Meanwhile, losses to the German workforce during the Second World War were compensated by large numbers of refugees and displaced persons from Eastern Europe and Eastern Germany: 8 million people arrived in West Germany between 1945 and 1974. In the 1950s, the Federal Republic commanded the largest reserve workforce in Europe in the 1950s (Wehler 2009, pp. 34–52). Due to the combined effects of the currency reform, the end of price controls, debt restructuring and Marshall Plan aid, German growth recovered, with growing shares of industrial production. New businesses were established, and unemployment rates fell significantly – the 'German economic miracle' began to set in. While German political and military power continued to be acutely constrained in the 1950s, Germany's economic performance within the European Communities flourished, with foreign trade becoming an engine of German growth. Between 1952 and 1958, the German gross national product (GNP) rose by more than half from around DM134 bn to DM198 bn (Bührer and Schröder 1992, pp. 175–6).

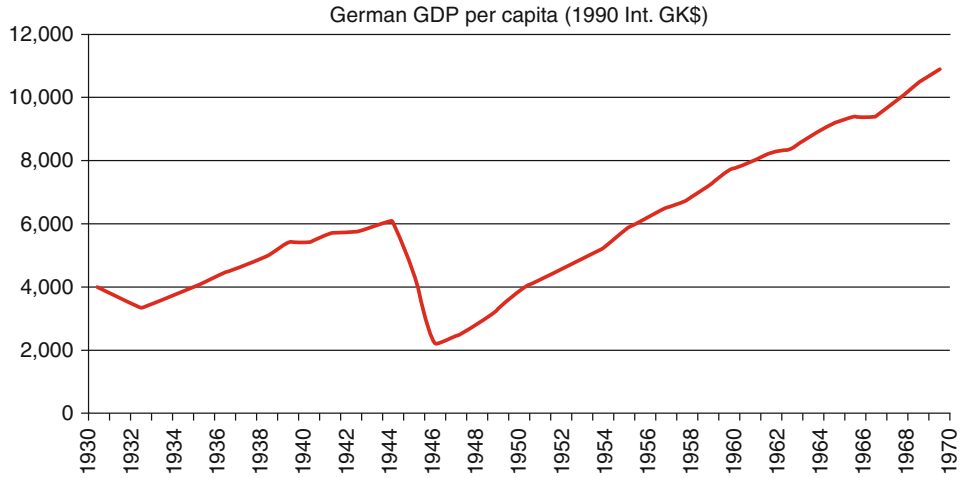
Figure 1 shows how strongly the German economy grew in the post-war years. The average annual growth rate was 9.5% between 1950 and 1955, and 8.5% from 1955 until 1960. The share of exports as a percentage of GNP rose from 9% in 1950 to 19% in 1960. By 1961, unemployment had gone down to less than 1% (Wehler 2009, pp. 54–5) and GDP per capita reached pre-war levels in the 1950s (Fig. 2) and increased further from thereon.

The structure of the German economy also adapted quickly. Western Germany began to export less raw material and semi-finished goods, but more industrialised products like machines and vehicles, with a higher value added. In the 1960s, the Federal Republic was back among the leading world economies (Bührer and Schröder 1992, pp. 175–6; Dernburg 1955; Wehler 2009).

It is Germany's rapid regaining of economic strength that explains why its European partners saw European integration as a key instrument to constrain Germany's economic and political



Germany's Historical Relationship with the European Union, Fig. 1 German GDP growth in the post-war period (Source: http://www.ggdc.net/Maddison/Historical_Statistics/horizontal-file_03-2007.xls)



Germany's Historical Relationship with the European Union, Fig. 2 German GDP per capita since 1930 (Source: http://www.ggdc.net/Maddison/Historical_Statistics/horizontal-file_03-2007.xls)

power in a cooperative and legally binding framework. Its Western neighbour France was especially afraid that a resurgent Germany would endanger French security, at the same time being aware that the growth of the German economy could not be stopped altogether, and was also beneficial to France's economic development. Figure 3 shows that, while Germany's GDP was lower than France's immediately after the Second World War, it grew more rapidly as of 1948.

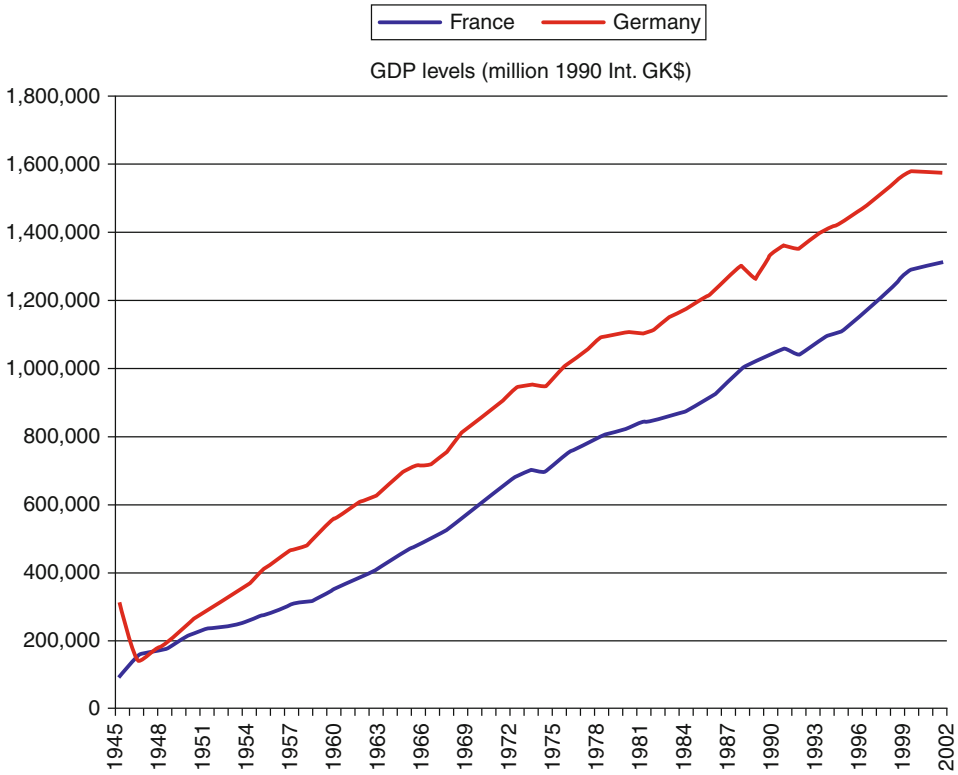
Germany and the Shaping of the European Order

It was hence only as part of the European integration project that Germany was accepted as an economic competitor without being perceived as a threat (Jeffery and Paterson 2003). Although being a 'tamed power' within Europe (Bulmer 2014, p. 1249), the country became a key shaper of the European economic order and its institutional structures. Both its rapid regaining of economic weight and its close cooperation with France on all major steps of integration explain how the defeated country could gain influence over the integration process so soon, in particular over the shaping of the single market, including competition policy, the move towards eastern enlargement and creation of the European Monetary Union.

The Single Market

Market liberalisation has always been a priority for Germany due to its strong export orientation, and the country has benefitted greatly from each step of economic integration in Europe. In the immediate post-war period, the German government was focused on institutionalising international trade at the West European level and thus extending its export markets, as it had limited access to export markets on its own due to its restricted sovereignty until 1955. The ECSC Treaty of 1951 established a common market for coal and steel among the six founding members and liberalised access to different production materials. The European Economic Community (EEC) later built on this structure and established a common market and a customs union, based on the free movement of persons, services, goods and capital. Moreover, the EEC treaty enshrined common policies like trade, agriculture and transport which is the first example of a Franco-German deal combining market integration and accompanying sectoral policies.

A particular priority from a German perspective was the Single European Market, which the Single European Act, signed in 1986, added to deepen the EEC. Chancellor Helmut Kohl had announced at the beginning of his chancellorship that he would promote the completion of the



Germany's Historical Relationship with the European Union, Fig. 3 German and French GDP from 1945 to 2002 (Source: www.ggdc.net/Maddison/Historical_Statistics/horizontal-file_03-2007.xls)

single market and fight protectionism within Europe, for instance by lowering agricultural subsidies (Kessler 2010, p. 124). This was a change in tone from a German government, as for decades Franco-German compromises had combined steps of market liberalisation or the creation of a European Competition policy based on the German model, on the one hand, and the introduction of redistributive policies (in particular agricultural and regional policy), on the other.

The programme completing the single market in 1992 met German interests very well, as it was first implemented in the manufacturing sector, in which Germany had been a strong actor and had managed to exert considerable influence on European standards (Jeffery and Paterson 2003, p. 66). While Germany, in contrast to France, has traditionally advocated a strong European competition policy (including state aid and merger control), there have also been cases in which the German government and the European Commission have been fiercely opposed over state aid to industry, public procurement procedures or the role of the regional state banks in providing subsidised finance to business. So Germany's continued arguments for a strong European Commission in the implementation of supranational competition policy have seen its relationship with the Directorate General for Competition Policy deteriorate over time.

One of the more contentious issues is the liberalisation of services, which conflicts with Germany's concern to shield parts of its services sector from international competition, and it has traditionally not been a driver of the norm- and agenda-setting in this sector. One major fear has been low-wage competition from central and eastern Europe, which became a politicised issue in the German domestic debate of the late 1990s when the effects of long-term liberalisation of the Single Market Programme began to penetrate deeply into the German economy. Unemployment

at the time was high and the economy of the reunited Germany was struggling. As a consequence, the enlargement of the EU to central and eastern European countries became a controversial issue in Germany, as for the first time since the start of European integration in the post-war period, distributional issues and Germany's net contribution to the EU became intensely discussed.

The German government, meanwhile, had promoted eastern enlargement early on, out of economic interests, a sense of historical and moral responsibility, and a strong interest in political stability in its eastern neighbourhood (Kessler 2010, p. 163). Eastern enlargement was hence a priority of the German Council Presidency in 1994. But the more eastern enlargement appeared as a competitive challenge to Germany, given its high wage and non-wage costs (Jeffery and Paterson 2003, p. 65), the more German policy-makers, especially from eastern German border regions, argued for transitional restrictions on the free movement of labour and on extensions of the single market to delay the anticipated negative effects. German industrial corporates, meanwhile, seized the opportunity of new markets and also potential new production sites and, for example, relocated part of German automotive production to central and eastern European countries. In contrast to other major European car producers, they found ways to profit from the lower labour-related costs and tax structures in some countries, so in sum, the eastern enlargement of the EU and the new member states' full integration into the single market turned out to be more of an opportunity than a threat for Germany due to its geographic proximity, its efforts to increase its own competitiveness and the pro-active seizure of opportunities in the new EU member states by German business. The anticipated negative effects on German labour markets never fully materialised.

A study by the Bertelsmann Foundation (2014) assesses the effects of 20 years of the single market, including the eight years following the EU enlargement of 2004 (1992 to 2012). It confirms previous studies that Germany benefitted largely from its integration into the single market and finds that it is among the countries most integrated into the European economy (followed by a number of small, open economies such as Denmark, Belgium and Austria). Germany and Denmark lead among all EU countries in terms of GDP per capita growth which the study finds are effects of the single market (see Table 1).

These beneficial growth effects of European integration, on top of peace and stability on the European continent, which were crucial for Germany to develop in the way that it could after 1945, need to be included in the discussion on the benefits and costs of European integration, which are all too often measured only by financial flows in the EU, in particular Germany's net contribution to the EU budget, which over the years, has averaged at 0.5% of German GDP.

Domestically, the defining moment for Germany's cost and benefit assessments of integration was German reunification in 1990, which led to a more fractious political landscape and

Germany's Historical Relationship with the European Union, Table 1 Average and cumulative gained income through European integration on the national level between 1992 and 2012

Position	Country	Average annual gained income since 1992 (€bn[a])	Cumulative gained income since 1992 (€bn[a])	Cumulative gained income since 1992 in relation to real GDP of 2012 in %
1	Germany	37.1	779	31.5
2	France	7.1	149	8.3
3	Italy	4.9	103	7.4
4	Spain	3.1	65	7.0
5	Denmark	2.7	57	27.2
6	Austria	2.3	48	17.5

Source: Prognos (2014), quoted in Bertelsmann (2014, p. 30)
[a]Real prices 2005; rounded

economic pressure. The German population increased by 16 million former GDR citizens to around 79.8 million people. In 1989, GDR growth rates and labour productivity were low, unemployment was high and the real economy lacked competitiveness. With reunification, the German government had to shoulder high costs in tackling the divide in economic structures, employment and income. The costs for structural reforms, solidarity funds and transfer payments from the west to the east of Germany consumed resources that otherwise might have been devoted to European integration. As a consequence, Germany was less willing to take on a great financial burden in the EU, in particular as the threat of the Cold War and the necessity to make Germany credible and a committed European country had lost importance. It started to argue for a reform of costly EU policies (agriculture and regional/cohesion policy). In the 1990s, German unity put the German economy under significant pressure, instead of turning Germany into an 'economic superpower' as some European states had feared (Kessler 2010, p. 139).

Economic and Monetary Union

Meanwhile, it was precisely the political momentum of Germany's reunification that finally pushed the member governments to actually agree on the creation of monetary union with the Maastricht Treaty of 1991, very much along the same logic of closely integrating Germany into European structures to contain its further growing economic and political weight.

In addition to the completion of the single market, the creation of the Economic and Monetary Union (EMU) has been a central priority of German policy-makers over the years. Over the decades, several attempts driven by France and Germany to stabilise exchange rates among the member states of the European Communities had eventually failed. Competitive devaluations posed a permanent challenge to Germany's export-based economy and distorted attempts to abolish trade barriers in the single market. Market integration and currency integration were hence seen as two closely interrelated projects from the German perspective. Germany and a few countries that had pegged their exchange rate to the Deutschmark (the so called ark bloc including e.g. Austria, the Netherlands and France) came under considerable economic pressure as their currencies appreciated over time, and there was a real risk in the 1980s that France, for example, could opt out of this arrangement to devalue. This would have pushed up the Deutschmark even further and would have hampered Germany's price competitiveness and hence its export performance considerably. So there were in fact strong economic reasons to push for a single currency to complement the single market, and German business lobbied strongly (see Collignon and Schwarzer 2005) for this project, the idea of which goes back decades, with the initial idea being presented to European Heads of State and the government of the European Communities with the Werner Plan in 1970.

In the Maastricht negotiations, Germany played a key role in shaping the setting-up of European Monetary Union, and managed to include many of its domestic structures, norms and policy principles at the EU level (Bulmer 2014; Jeffery and Paterson 2003, pp. 61–2), such as the independence and objectives of the European Central Bank. There was no unanimity on the model to be chosen for the European Central Bank. According to the Anglo-French model, central banks pursue not only monetary policy, but also other economic goals. Meanwhile, the German model of the Bundesbank suggested a politically independent central bank whose sole goal is price stability. During the negotiations, it became clear that Germany would only accept a European currency if it were as strong as the Deutschmark and run by a Central Bank which would follow the same principles as the Bundesbank. Eventually, the German central model was chosen over the Anglo-French model against the opposition of a number of governments, but with considerable backing from the presidents of national central banks.

Germany's negotiating position was based on firm assumptions about how the European economy and monetary system should work, which are deeply anchored in two economic paradigms which have dominated German post-war economic thinking: the neo-classical and the ordo-

liberal schools of thought. According to the ordo-liberal tradition, a strong state is needed to ensure a competitive economy, with the role of the state being restricted to regulation and setting a strong legal framework. A monetary union should be based on monetary stability, granted by an independent central bank, while public spending should be low and no bail-out option should be provided for governments that generate unsustainable public finances. As the neo-classical paradigm suggests, growth should be achieved through competitiveness and supply-side reforms, not through discretionary fiscal stabilisation or supply-side measures, which in French economic thinking, for example, are important to support domestic consumption.

While the Maastricht Treaty contained a few basic principles for the coordination of national economic and budgetary policies, the German government put forward an initiative to strengthen the rules-based coordination approach, which was later called the Stability and Growth Pact (SGP) and was agreed upon in 1997. As concessions to France, a growth element was included in the deal, and the Eurogroup (initially named Euro-X and introduced without any Treaty base) was created to provide a political forum for the finance ministers. This package deal is one example of a way to strike a balance between Germany's concern to establish rules-based coordination as firmly as possible, while consecutive French governments sought ways to create space for flexibility, discretion and a more political approach to the governance of the single currency and the European economy. Contrasting perspectives on how the European economy would and should work also surfaced in the discussion about how to select the members of the currency union. The economist camp (Germany, the Netherlands and Austria) demanded that economic and fiscal convergence precede monetary integration, so that the costs of adjustment would be solely shouldered by the converging countries. Monetary union should thus be the 'coronation' of economic integration. The monetarist camp of France, Belgium and Luxembourg meanwhile was convinced that monetary integration would lead to economic convergence and trigger political integration. So the application of convergence criteria should be less of a priority.

When the euro was introduced in 1999, a monetary union started that indeed carried a strong German imprint. However, the fundamental conflict between opposing views on economic and monetary policy, represented by Germany and France, was never resolved. Firstly, although Maastricht was sold in Germany as modelling the euro on the Deutschmark's success, it did represent a careful balance between the opposing positions described above. Secondly, over time, there had to be a readjustment and Germany drove one of the most discussed of them: in 2003, Germany and France violated the SGP and pushed for its subsequent reform, which led to more flexibility to account for the cyclical situation of the member states. While the adjustment was widely seen as economically reasonable, it was also interpreted as a change in Realpolitik, with Germany and France as major powers pushing through their interests against European rules. Indeed, over time, national interests and assertiveness became more evident in Germany policy practice (Bulmer 2014, p. 1249).

Also in the following years, and in particular during the crises that hit the euro area from 2007/08 onwards, the opposing visions of governing a single currency and the European economy resurfaced. Germany has been highly committed to keeping the currency union together, while upholding principles of monetary stability, fiscal conservatism and a rules-based approach to governing the euro area. On the EU level, German policy-makers enforced a strict austerity policy against the resistance of the majority of eurozone states which shaped the notion of Germany's 'reluctant and contested hegemony' (Bulmer and Paterson 2013, p. 1401) (for more information on Germany's role in the Eurozone crisis see Schwarzer (2014).

The Institutional Setup of the European Union

As a result of its two world war experiences and inspired by its own federal political system,

Germany since early on in the process of European integration, has promoted strong supranational structures. It has traditionally advocated a strong European Parliament and a strong European Commission in particular to oversee the governance of the single market, including Competition policy, as well as a politically and legally independent European Central Bank. It has supported the introduction and expansion of majority voting in the Council, and except for Chancellor Schröder during the negotiations of the Nice Treaty, it has never questioned that Germany, France, Italy and the UK have equal voting weights, despite significant differences in population numbers.

As early as in the 1950s, Chancellor Adenauer advocated the establishment of a political union, whereas the French government lobbied for a 'Europe of States' based on intergovernmental cooperation (e.g. Fouchet Plan). This trend continued with later governments and confirms that Germany and France, for a long time, have had different ideas on the final goal of European integration. During the Maastricht negotiations, some of the strongest voices for a political union to accompany monetary union came from Germany, including government representatives and the influential President of the European Central Bank. In 1998, then-Foreign Minister Joschka Fischer started another debate on the future political structures of the EU, promoting elements of a federalist Europe and a European Constitutional Treaty.

Despite this strong tradition of supporting supranational institutions and decision-making, the German position as to how far Europe should be integrated has evolved over time. First of all, the Constitutional Court in its Maastricht ruling characterised EU as an 'association of states', while it also strengthened the role of the German Länder and of the German Parliament. Germany has since strongly emphasised the subsidiarity principle as a guideline for further decisions on European integration. Moreover, criticism of the European Commission, changing perspectives on Germany's financial contribution and generally more realism about the future of the EU have appeared in German discourse.

Foreign and Security Policy

Germany has traditionally been in favour of closer cooperation in European foreign and security policy as well as defence, but, in contrast to France, it was absent from international military missions in the context of NATO or European foreign policy for a long time. This was mainly due to Germany's tradition as a civilian power and very critical public opinion concerning military deployment abroad. Germany, however, has been involved in humanitarian missions and has contributed financial aid and equipment.

The fact that Germany became involved in the Kosovo war in 1999 and took part in the Operation Enduring Freedom as well as the International Security Assistance Force (ISAF) NATO mission in Afghanistan is widely considered a turning point in German foreign policy, as Germany no longer reflexively rules out participation in large-scale military interventions.

The flip side of this 'normalisation' process, however, was that Germany took decisions independently of its key allies, e.g. not to take part in military missions in Iraq or to abstain from the vote in the UN Security Council on Resolution 1973 which imposed a no-fly zone over Libya. This made the EU appear divided and raised concerns among partners like the USA, France or the UK that Germany might not be a reliable partner in foreign and security policy.

In fact, international expectations and the need to follow its military allies and guarantors of security are no longer the predominant driving force in German foreign policy. Since the end of the Cold War and German reunification, foreign policy has become more politicised and contested in the domestic arena. One reason is that with the dissolution of the Cold War order a more heterogeneous spectrum of foreign policy orientations has developed, as policy-makers gained more leeway because German foreign policy was less determined by Germany's geopolitical position. Moreover, the German Bundestag became a stronger actor. Since 1994 the German parliament has had the final say over military deployments, which constrains government decisions on Bundeswehr missions.

Conclusion: Continuity and Chance in Germany's Role in the EU

Since the creation of the European Communities in the 1950s, German policy-makers have continually emphasised Germany's European vocation and promoted further European integration. Germany is seen by many scholars as 'the most consistently pro-integrationist member state' (Bulmer 2014, p. 1248). German governments, for a long time, did not pursue their policy goals unilaterally or articulate their interests in national terms. Instead, they framed priorities as shared interests and pursued them in alliance with their European partners, mostly France. This 'leadership avoidance reflex' (Jeffery and Paterson 2003, p. 61) has been especially distinctive in security terms, where Germany has been a 'policy taker', relying on the USA as well as other EU member states like France or the UK to take the lead. German leadership on the EU matters has been more pronounced in economic and monetary issues, in particular since Germany moved to centre stage as a key manager of the crises in the euro area since 2007/08. Germany has often been characterised as an 'economic giant' and 'political dwarf' (Bulmer and Paterson 2013, p. 1388).

Until German reunification in 1990 Germany's European diplomacy was characterised by a strong and stable public and political consensus in favour of more European integration. This pro-European discourse went hand in hand with support for concrete initiatives concerning institutional reforms, foreign policy cooperation, enlargement or completion of the single market. With the Treaty of Maastricht, the EU took on several new challenges and pursued institutional reforms, the implementation of economic and monetary union and negotiations on Eastern enlargement in the wake of the demise of the Soviet Union.

Since 1990, Germany's role in the EU has undergone far-reaching changes (Jeffery and Paterson 2003, p. 63). Germany's role as 'economic giant' and net payer of the EU became more contested within Germany. German policy-makers increasingly emphasised cost–benefit calculations in their European policies, focusing more on German interests and less on the expectations of other EU member states. This trend first became evident in the mid-1990s under Helmut Kohl's leadership and increased further in the late 1990s with the Red–Green coalition government, led by Chancellor Gerhard Schröder (Jeffery and Paterson 2003, p. 68; Kessler 2010, p. 124). Kohl and Schröder attempted to lower the German share of the EU budget and to increase German influence in decision-making. Another example of this 'new' interest-driven governance was the breach of the Stability and Growth Pact in 2003. The latest case of more unilateral politics has been eurozone crisis management, where Germany took on a leadership role in shaping crisis responses in accordance with German interests and models, at times against strong opposition from other EU member states. Bulmer even talks about a 'departure from [Germany's] past role as "benign hegemon"' (Bulmer 2014, p. 1249).

Along with the tough economic challenges in the country the political landscape became more fractious, with the Socialists (PDS) joining the Bundestag and shaping the national discourse. Another aspect in the wake of reunification was the increasing territorialisation of politics, with the German Länder asking for more leeway and participation rights in domestic and European policy-making, which were eventually granted to them by the Maastricht ruling of the Constitutional Court. This ruling also enhanced the position of the Bundestag in the context of an increasing shift of sovereignty to the EU level, adding to the contestation of European policy-making in the German domestic discourse.

The process of increasing articulation of national interests – also in monetary terms – and the normalisation of German foreign policy showed that German policy-makers have become more self-confident in their actions, more explicit in their rhetoric, less reflexive in choosing multilateralism when promoting their national interests and less worried about disappointing their international partners (Oppermann 2012). Chancellor Gerhard Schröder announced in 1998 that 'Germany standing up for its national interests will be

just as natural as France or Britain standing up for theirs' (Paterson 2011, p. 65). These changes, combined with the new post-Cold War realities in the EU, have brought back fears of German hegemony in some parts of Europe. France and the UK were sceptical of German reunification because they expected a bigger Germany to expand its political power and claim more leadership within the EU (Rudzio 2011, p. 20). In particular, France was also very hesitant about Eastern enlargement in 2004 because it considered the accession of Poland and other central and eastern European states as a further shift of power towards Germany in political as well as economic terms, moving the country further towards the centre of the EU and endangering the French leadership position within the EU (Lepenies 2014). This sensitivity to German power is one of the reasons why German governments have essentially sustained their pro-European visions and adherence to multilateral policy-making. However, Germany's role in the recent management of the sovereign debt crisis in the euro area has given rise to further questions and criticism (Germany in the euro area crisis, 2014).

See Also

- ▶ Euro
- ▶ European Central Bank And Monetary Policy In The Euro Area
- ▶ European Monetary Integration
- ▶ European Monetary Union
- ▶ European Union Single Market: Design And Development
- ▶ European Union Single Market: Economic Impact
- ▶ Germany, Economics in (20th Century)
- ▶ Germany in the Euro Area Crisis
- ▶ German Reunification, Economics of

Bibliography

Bertelsmann Foundation. 2014. *20 Jahre Binnenmarkt: Wachstumseffekte der zunehmenden europäischen Integration*. https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/20_Jahre_Binnenmarkt-de_NW.pdf.

Bührer, W., and H.-J. Schröder. 1992. Germany's economic revival in the 1950s: The foreign policy perspective. In *Great Britain, France, Germany and Italy and the origins of the EEC, 1952–1957*, ed. E. Di Nolfo, 174–196. Berlin/New York: Walter de Gruyter.

Bulmer, S. 2014. Germany and the eurozone crisis: Between hegemony and domestic politics. *West European Politics* 37(6): 1244–1263.

Bulmer, S., and W. Paterson. 2013. Germany as the EU's reluctant hegemon? Of economic strength and political constraints. *Journal of European Public Policy* 20(10): 1387–1405.

Dernburg, H.J. 1955. Rearmament and the German economy. *Foreign Affairs* 33(4): 648–662.

Federal Ministry of Defence. 2013. *The history of the Bundeswehr*. http://www.bmvg.de.

Henderson, D.R. 2007. German economic miracle. In *The concise encyclopedia of economics*. http://www.econlib.org/library/Enc/GermanEconomicMiracle.html.

Jeffery, C., and W. Paterson. 2003. Germany and European integration: A shifting of tectonic plates. *West European Politics* 26(4): 59–75.

Kessler, U. 2010. Deutsche Europapolitik unter Helmut Kohl: Europäische Integration als 'kategorischer Imperativ'? In *Deutsche Europapolitik: Von Adenauer bis Merkel*, ed. G. Müller-Brandeck-Bocquet, C. Schukraft, N. Leuchtweis, and U. Kessler, 119–171. Wiesbaden: VS Verlag für Sozialwissenschaften.

Lepenies, W. 2014. Frankreichs Furcht vor dem Fünften Reich. *Die Welt*, November 11. http://www.welt.de/kultur/article134200796/Frankreichs-Furcht-vor-dem-Fuenften-Reich.html.

Oppermann, K. 2012. National role conceptions, domestic constraints and the new 'normalcy' in German foreign policy: The eurozone crisis, Libya and beyond. *German Politics* 21(4): 502–519.

Paterson, W.E. 2010. Does Germany still have a European vocation? *German Politics* 19(1): 41–52.

Paterson, W.E. 2011. The reluctant hegemon? Germany moves centre stage in the European Union. *Journal of Common Market Studies* 49 (Annual Review): 57–75.

Rudzio, W. 2011. *Das politische System der Bundesrepublik Deutschland*. Wiesbaden: VS Verlag.

Ritschl, A. 2012. Germany, Greece and the Marshall Plan. *The Economist*, June 15. http://www.economist.com/blogs/freeexchange/2012/06/economic-history.

Schwarzer, D., and S. Collignon. 2005. Unternehmen und Banken auf dem Weg zur Währungsunion: Die "Association for the Monetary Union of Europe" als Motor eines transnationalen Konsenses. In *Interessenpolitik in Europa*, ed. B. Kohler-Koch and R. Eising, 203–226. Baden-Baden: Nomos.

Wehler, H.-U. 2009. *Deutsche Gesellschaftsgeschichte 1949–1990*. Bundeszentrale für politische Bildung.

Gerschenkron, Alexander (1904–1978)

Albert Fishlow

Keywords

Backwardness; Capital accumulation; Economic history; Gerschenkron, A.; Industrial revolution; Industrialization; Laspeyres index; Latecomer development; Modernity; Paasche index; Stages of growth; Structural change

JEL Classifications

B31

Gerschenkron was born in Odessa in 1904 and died in Cambridge, Massachusetts, in 1978. He left Russia in 1920 and settled in Austria. In 1938, a decade after receiving the degree of *doctor rerum politicarum* from the University of Vienna, he emigrated to the United States and spent the next six years at Berkeley. After a short period at the Federal Reserve Board, he went to Harvard in 1948 to teach both economic history and Soviet studies. His passion for the former dominated, and he flourished there as the doyen of economic history in the United States. He influenced a generation of Harvard economists through his required graduate course in economic history and attracted several to his seminar and the field. His erudition and breadth were legendary, and defined an indelible, if unattainable, standard of scholarship for his colleagues and students.

Gerschenkron's principal contribution to economics was the elaboration of a model of latecomer economic development. Its central hypothesis is the positive role of relative economic backwardness in inducing systematic substitution for supposed prerequisites for industrial growth. State intervention could, and did, compensate for the inadequate supplies of capital, skilled labour, entrepreneurship and technological capacity found in follower countries. Thus the German institutional innovation of the 'great banks' provided access to needed capital for industrialization, even while greater Russian backwardness required a larger and more direct state role.

Gerschenkron's analysis is consciously anti-Marxian: it rejected the English Industrial Revolution as the normal pattern of economic development, and deprived the original accumulation of capital of much of its conceptual force. Elements of modernity and backwardness could survive side by side, and did in a systematic way. Apparent disadvantageous initial conditions of access to capital could be overcome. Success was rewarded with proportionately more rapid growth, signalled by a decisive spurt in industrial expansion.

This model, first presented in 1952 in an essay entitled *Economic Backwardness in Historical Perspective* (reprinted in 1962), underlay Gerschenkron's extensive research into the specific developmental experiences of Russia, Germany, France, Italy, Austria and Bulgaria. Out of those historical studies emerged a comparative, all-encompassing European picture. 'In this fashion, the industrial history of Europe is conceived as a unified and yet graduated pattern' (Gerschenkron 1962, p. 1). In turn, his hypotheses became progressively more precise. They may be summarized as follows:

1. Relative backwardness creates a tension between the promise of economic development, as achieved elsewhere, and the reality of stagnancy. Such a tension motivates institutional innovation and promotes locally appropriate substitution for the absent preconditions of growth.
2. The greater the degree of backwardness, the more interventionist was the successful channelling of capital and entrepreneurial guidance to nascent industries. Also, the more coercive and comprehensive were the measures to reduce domestic consumption.
3. The more backward the economy, the more likely were: an emphasis upon producers'

goods rather than consumers' goods; use of capital-intensive rather than labour-intensive methods of production; emergence of larger rather than smaller units of both plant and enterprise; dependence upon borrowed, advanced technology rather than indigenous techniques.

4. The more backward the country, the less likely was the agricultural sector to provide a growing market to industry through rising productivity, and the more unbalanced the resulting productive structure of the economy.

The considerable and continuing appeal of the Gerschenkron model derives from its logical and consistent ordering of the process of European development, the conditional nature of its predictions and its generalizability to the experience of the late latecomers of the present Third World. His formulation rises above other theories which emphasize stages of growth both because of its attention to historical detail and its insistence upon the special attributes of latecomer development that cause differential evolution. In Gerschenkron's own hands, his propositions afforded an opportunity to blend ideology, institutions and the historical experience of industrialization, especially that of Russia, in a dazzling fashion. For others, his approach has proved a useful starting point for the discussion of non-European latecomers, including Japan and the newly industrializing countries.

The model is, of course, not without its limitations. History, even of Europe alone, does not in every detail bear easily the weight of such a grand design. In other parts of the world, as might be expected from a concept rooted in the special features of the historical European experience, larger amendments are frequently required. And somewhat surprisingly, in view of Gerschenkron's own path-breaking essay in political economy, *Bread and Democracy in Germany* (1943), there is too little attention to the domestic classes and groups whose interests the interventionist state must adequately incorporate if it is to play the central role required. Backwardness too easily becomes an alternative, technologically rooted explanation, distracting attention from the state rather than focusing upon its opportunities and constraints.

Still, the concept of relative backwardness, and Gerschenkron's always insightful and rich elaborations in so many national contexts, represent a brilliant and original contribution to economic history for which he is justly celebrated. It is not the only one. The 'Gerschenkron effect', arising from the difference between calculated Paasche and Laspeyres indexes of Soviet machinery output (1951), also commemorates him. Current price weights will tend to underestimate the extent of growth because prices and quantities are negatively correlated, just as base year weights exaggerate it. The larger is the difference between the alternatively constructed quantity indexes, the greater is the degree of structural change. Again, divergence rather than uniformity is the source of useful information about historical processes.

Alexander Gerschenkron has few peers, past or present, in his command of comparative economic history. Scholarly interest in contemporary economic development has brought him an increasing following. His insights thus continue to influence a new generation of scholars and guarantee him a central place in any assessment of the evolution of the discipline of economic history.

Selected Works

1943. *Bread and democracy in Germany.* Berkeley: University of California Press.
1951. *A dollar index of soviet machinery output*. Santa Monica: Rand Corporation.
1962. *Economic backwardness in historical perspective*. Cambridge, MA: Harvard University Press.
1968. *Continuity in history and other essays.* Cambridge, MA: Harvard University Press.
1970. *Europe in the Russian mirror.* London: Cambridge University Press.
1977. *An economic spurt that failed.* Princeton: Princeton University Press.

Gervaise, Isaac (*fl.* 1680–1720)

Peter Groenewegen

Keywords
Free trade; Gervaise, I.; Law, J.; Specie-flow mechanism

JEL Classifications
B31

Merchant and economist of French Huguenot extraction. Gervaise was born in the second half of the 17th century, probably in Paris, and migrated with his family to London in 1681. With his father he was associated with the Royal Lustring Company (1688–1720) engaged in the manufacture of a fine, light, black, glossy silk under patent granted by parliament. Ironically, the year the company lost its charter saw the publication of Gervaise's 34-page pamphlet, *The System or Theory of the Trade of the World* with its attack on exclusive companies. Foxwell (1940, p. 167) described it as 'one of the earliest formal systems of political economy . . . stating one of the most forcible practical arguments for free trade'. Quite unlike much contemporary writing on trade, Gervaise's pamphlet is tersely written and especially noted for its peculiar terminology and highly abstract argument. Gervaise is presumed to have died in London by 1739.

The real discoverer of Gervaise's work, Viner (1937, pp. 79–80), has described it as 'an elaborate and close reasoned exposition of the nature of international equilibrium and of the self-regulating mechanism whereby specie obtained its "natural" or proper international distribution'. The novelty of Gervaise's treatment of the specie mechanism is his emphasis on the role of income rather than prices, in strong contrast with subsequent treatments by Cantillon, Vanderlint and Hume. The starting point for the analysis is the proposition that the equilibrium bullion stock of any nation is proportioned to its output in terms of labour and that such a stock also maintains the balance between consumption and production, exports and imports.

Excess bullion breaks these balances by raising consumption and reducing production, thereby lowering exports and raising imports, hence bullion will be exported and the balances will be resorted. An inadequate bullion stock leads to specie inflow by raising production relative to consumption, and exports to imports. Gervaise treats credit as if it were bullion; oversupply or deficiency is self-correcting via the balance of trade, though the adjustment process with credit is more rapid through its additional income effects of interest payment to suppliers of credit, whom he sees as consumers rather than producers. Hence 'credit is of pernicious consequences to that Nation that uses or encourages it beyond nature' (Gervaise 1720, p. 14) – a comment perhaps not unrelated to contemporary developments with Law's system in France. War, capital consumption or export, and restrictions on trade may prevent or postpone attainment of monetary equilibrium. For this reason, and for the resource misallocation potential flowing from encouragement of specific manufactures through companies, laws or taxes on imports, Gervaise (1720, pp. 17–18) concluded that 'Trade is never in a better condition, than when it's natural and free'. Gervaise also pointed out that the 'natural proportion' of bullion for specific countries was influenced by their situation, particularly as regards proximity to water transport, and that implementing policies of debasing the currency had similar effects on trade and the balance of consumption and production as credit oversupply. Although less elegantly written than Hume's later account of the specie mechanism, the emphasis on adjustment through income rather than price effects, though not always clearly explained, makes Gervaise's short and penetrating contribution to the subject more modern than most of its successors in the century and a half which followed.

Selected Works

1720. *The system or theory of the trade of the world*. Reprinted with a biographical

introduction by J.M. Letiche and a foreword by J. Viner. Baltimore: Johns Hopkins Press, 1954.

Bibliography

Foxwell, H.S. 1940. Comment reproduced in *Catalogue of the Kress Library of Business and Economics.* Boston: Baker Library, Harvard Graduate School of Business Administration

Viner, J. 1937. *Studies in the theory of international trade*. New York: Harper and Brothers.

Gesell, Silvio (1862–1930)

Victoria Chick

Keywords

Depressions; Gesell S.; Hoarding; Land nationalization; Stamped money principle

JEL Classifications

B31

Gesell was born in Germany but emigrated to the Argentine in 1886, where he was so successful as an importer that he retired to Switzerland in 1900 to farm and to continue to write. The 'retirement' included a return to the Argentine to manage his late brother's business and an involvement in Bavarian politics at their most chaotic. As a deposed minister of finance he was tried for high treason, and acquitted.

His prolific writing began in Argentina, provoked by the economic chaos of the late 1880s there. But his fame rests on *The Natural Economic Order,* originally published in two parts in 1906 and 1911. It was translated into English in 1929. Rent-free land and interest-free money characterize that Order. Land would be nationalized, its owners compensated by the issue of state bonds. Through the device of stamped money, which would remain current only if a stamp, obtained at a cost set by government, was regularly affixed, the rate of interest on these bonds and other lending instruments would eventually be driven to zero. With no income diverted to rent or interest the worker would receive the full value of his output. Mothers were to receive income from annuities based on the nationalized land, since their 'output', the population, was the source of demand for land and hence rent.

Gesell attributed depressions to inadequate investment and the latter to the fall in the expected rate of return as investment continued, coupled with a money rate of interest which was prevented from falling by the alternative opportunity of hoarding. This analysis substantially anticipates Keynes's (1936), as Keynes amply acknowledges (pp. 353–8). Gesell suggested adjusting the stamp duty on money to force down the rate of interest.

The stamped money principle was three times applied on a local scale in the 1930s: in Bavaria, in the Austrian Tyrol and in Alberta, Canada. In each case the scheme successfully raised demand and employment, but the money was soon banned by the authorities.

Though theoretical inadequacies and practical difficulties are claimed against Gesell's theory, its *aim* is probably more responsible for its eclipse. But it lives on furtively, below the surface, in the underworlds of Keynes's *General Theory* and Fisher's *Booms and Depressions.*

Selected Works

1891a. *Currency reform as a bridge to the social state*. Buenos Aires. Trans. P. *Pye, typescript*, 1951.

1891b. *Nervus Rerum: Continuation of 'Currency reform as a bridge to the social state'*. Buenos Aires. Trans. P. Pye, typescript, 1951.

1929. *The natural economic order: A plan to secure an uninterrupted exchange of the products of labour.* Trans. P. Pye, Berlin: Neo-Verlag. Online. Available at http://www.systemfehler.de/en/neo/. Accessed 10 Nov 2006.

Bibliography

Fisher, I. 1933. *Booms and depressions*, Appendix 7. London: George Allen & Unwin.
Gaitskell, H. 1969. *Four monetary heretics*. Christchurch: Lyn Christie & Son.
Keynes, J.M. 1936. *The general theory of employment, interest and money*. London: Macmillan.
Wise, L. n.d. *Silvio Gesell*. London: Holborn Publishing Co.

Ghettoes

Robert A. Margo

Abstract

A ghetto is an area of a city in which a minority group is highly segregated and kept there by social, legal, or political forces. This article describes the history of Jewish ghettos of medieval Europe through the Nazi period. The African-American experience is also described, along with the effects of segregation on racial differences in economic outcomes. Examples of ghettos in Japan (the Burakumin), Australia (aborigines), and South Africa (apartheid) are mentioned, along with the related concept of an ethnic enclave.

Keywords

Apartheid; Burakumin; Ghettoes; Housing markets; Human capital; Immigration; Internal migration; Neighbours and neighbourhoods; Racial discrimination; Racial segregation; Restrictive covenant; Urban segregation

JEL Classifications

N90; N92; N93; N94; J15

A ghetto is a specific area of a city in which a racial, ethnic, or other minority group constitutes the overwhelming majority of residents, and which is maintained as such by social, political, or legal pressures.

In early medieval Europe local authorities granted Jews living quarters in exchange for services as traders, moneylenders, or tax collectors. With the Crusades, these voluntary arrangements turned compulsory, first in Spain and Portugal and later in Germany. The Jews of early 15th Venice were forced to live in the *ghetto Nuevo,* an abandoned foundry located on an island isolated from the general population (Curiel et al. 1990). Walls were erected and guards stood watch by the foundry gates. By day Jews could continue to conduct business but they were required to wear special clothing and were subject to punishment if they were found outside the ghetto at night. Within the walls of the typical Jewish ghetto an entirely separate world of institutions – courts, educational and cultural institutions, retail establishments, and recreational facilities – emerged. Overcrowding, however, was the norm because geographic expansion at the periphery of the ghetto was severely limited, if it existed at all. By the 19th century the extant Jewish ghettos were seen as outmoded relics of an earlier age. In continental Europe the last ghetto to be abolished was the Roman ghetto in the late 19th century. The so-called Pale of Settlement in the far West of Russia survived until the October Revolution of 1917. A few decades later, Jewish ghettos were revived in German-occupied Europe with a vengeance by the Nazis, who used them primarily as holding pens prior to extermination (Browning 1986).

In the early 20th century social scientists began to refer loosely to any area of a city in which a particular racial, ethnic, or religious group was very highly segregated as a 'ghetto' (Wirth 1928). Later in the century the usage spread to other types of segregation (such as in the phrase 'pink ghettos', referring to the concentration of women in certain occupations). Numerical indices were devised, primarily by sociologists, to measure the extent of segregation (Massey and Denton 1993; Echenique and Fryer 2004).

Wacquant (2004) argues that the following are necessary, if not always sufficient, characteristics of a ghetto in the looser sense. First, the minority group must be readily identifiable. Second, the group must be kept physically isolated by the majority, which has the power to enforce the

isolation. Third, institutions emerge within the ghetto to provide services that ghetto residents cannot otherwise obtain from the outside world. A distinct ghetto culture may also emerge, elements of which might persist long after the specific mechanisms that kept the affected group 'in its place' are abolished.

The word 'ghetto' has often been used to describe the experience of African Americans in urban areas in the 20th century (Drake and Clayton 1945; Clark 1965; Osofsky 1971; Hirsch 1983; Massey and Denton 1993; Cutler et al. 1999). In the late 19th century African Americans were concentrated in the rural southern United States, where they faced very high levels of racial discrimination. In the northern United States the African-American population was a small share of the total and of the urban population; and only in one city (Norfolk, Virginia) was the extent of racial segregation great enough to plausibly justify claims that a ghetto existed (Cutler et al. 1999).

Although a steady trickle of African Americans to urban areas had occurred since the end of the civil war, the trickle became a flood during the First World War. European immigration from Europe was largely cut off and, buoyed by wartime demand, urban manufacturers turned to southern blacks as a new labour supply. By 1940, virtually all major cities in the North had black ghettos, some known by specific names – 'Harlem' in New York, 'Bronzeville' in Chicago. Unlike the Jewish ghettos of the medieval Europe, the boundaries of the black ghettos were not set in stone, and whites living nearby adopted a variety of tactics, legal and otherwise, to limit the ghetto's geographic expansion. One example was the *restrictive covenant,* a clause contained in a deed of sale that enjoined a white owner from selling to a black family. Such covenants were in widespread use until rendered unenforceable by the United States Supreme Court in 1948. the Second World War had a much greater effect on black migration from the rural South than the First World War had, and segregation levels continued to increase, reaching 'staggering levels' by 1970 (Cutler et al. 1999, p. 470). Since 1970 racial segregation in American cities has declined, although levels remain relatively high compared with the late 19th century.

Economists have considered how ghettos affect economic and social outcomes of African Americans. In the model developed by Cutler and Glaeser (1997), there are three groups: skilled whites, skilled blacks, and unskilled blacks. The groups occupy a city which is divided into three neighbourhoods fixed in geographic size. Economic decisions are made by parents, and parental utility is a positive function of their children's human capital, which itself depends on the human capital of the parent and the average level of human capital in the community where the household resides. Utility is a negative function of housing costs and, in addition, blacks (whites) must pay a fixed entry cost to live in a white (black) neighbourhood.

Cutler and Glaeser consider an initial equilibrium in which whites live in one neighbourhood while skilled and unskilled blacks are distributed across the remaining two neighbourhoods. An increase in the cost to blacks of entering the white neighbourhood reduces the number of skilled blacks living in the white neighbourhood. White utility increases because housing costs fall and average skill levels are unaffected. As more skilled blacks choose to live in one of the black neighbourhoods, housing costs in the neighbourhood rise, which hurts unskilled blacks living there. However, average human capital in the neighbourhood increases, which benefits the human capital production of the children of unskilled blacks. The net effect on the location decisions of unskilled blacks is indeterminate but, depending on the overall effect, it is possible for welfare of unskilled blacks to decrease.

Cutler and Glaeser conduct an empirical analysis of the relationship between segregation and economic outcomes for African Americans, using census data for 1990. They show that an increase in residential segregation reduces educational attainment and income and increases the incidence of motherhood at younger ages. Collins and Margo (2000) replicate the empirical analysis for earlier census years, finding that black ghettos turned increasingly bad on all fronts after 1970. Collins and Margo (2003) study the evolution of racial differences in the values of owner- occupied housing since 1940. They demonstrate that, in the

country as a whole, the racial gap in housing values was narrowing prior to 1970. However, in central cities the racial gap widened in the 1970s, and the extent of widening was greater in cities that were initially heavily segregated in 1970.

If high levels of racial segregation turned increasingly bad for African Americans in central cities after 1970, there was no shortage of possible causes. Massey and Denton (1993) emphasize deindustrialization, whereas Wilson (1987) hypothesizes that decreases in housing market discrimination enabled middle- and upper-class Americans to leave central cities ghettos, thereby removing effective role models from the ghetto. Although the United States has a long (and terrible) history of race-related civil disturbances, the number and geographic extent of those occurring in the 1960s were unprecedented. Collins and Margo (2004a, b) demonstrate that the occurrence of a severe riot between 1960 and 1970 was associated with deteriorating employment, earnings, and housing values for urban blacks, an association that persisted at least through the 1970s.

European Jews and African Americans are far from the only groups whose histories include ghettos or ghetto-like conditions. One such group is the Burakumin of Japan. The Burakumin were descendants of the lowest of four castes that existed during the feudal era of Japanese history. Viewed as 'untouchable' by the religious majority (Buddhists and Shinto), the Burakumin were faced with similar constraints to those imposed on Jews, such as restrictions on intermarriage and geographic mobility. Like Jews, the Burakumin had to wear special clothing and behave in a deferential manner towards the majority. Although 'emancipated' by edict in 1871, a combination of formal and informal institutions nevertheless resulted in the formation of numerous impoverished and crime-ridden ghettos of Burakumin throughout Japanese cities that have persisted (DeVos and Wagatsuma 1966; Hane 1982). Brock (1993) describes the emergence, persistence, and functioning of three 'outback' ghettos (Poonindle, Koonidule, and Nepabunna) of Aboriginal people in Australia. Residential segregation was a key aspect of apartheid in South Africa as well as other colonial regimes (Abu-Lughod 1980; Western 1981). Large-scale international migrations have frequently resulted in ghetto-like conditions in the receiving countries (see Wirth 1928, for a classic discussion). Such 'ethnic enclaves' may acquire names – 'Chinatown' (see Zhou 1992) or 'Little Italy' – but are fundamentally different because the newcomers eventually assimilate into the broader society. The techniques developed for measuring racial segregation in the United States are now being routinely applied to other countries; see, for example, Peach (1996), and also Johnston et al. (2002), who shows that overall levels of ethnic segregation in British cities around 1990 were considerably lower than levels of racial segregation in American cities.

See Also

- ▶ Housing Policy in the United States
- ▶ Residential Segregation

Bibliography

Abu-Lughod, J. 1980. *Rabat: Urban apartheid in Morocco*. Princeton: Princeton University Press.

Brock, P. 1993. *Outback ghettos: A history of aboriginal institutionalism and survival*. New York: Cambridge University Press.

Browning, C. 1986. Nazi ghettoization policy in Poland, 1939–1941. *Central European History* 19: 343–368.

Clark, K. 1965. *Dark ghetto: Dilemmas of social power*. New York: Harper.

Collins, W., and R. Margo. 2000. Residential segregation and socioeconomic outcomes: When did ghettos go bad? *Economics Letters* 69: 239–243.

Collins, W., and R. Margo. 2003. Race and the value of owner-occupied housing, 1940–1990. *Regional Science and Urban Economics* 33: 255–286.

Collins, W., and R. Margo. 2004a. The labor market effects of the 1960s riots. In *Brookings-Wharton papers on urban affairs 2004*, ed. W. Gale and J. Pack. Washington, DC: Brookings Press.

Collins, W., and R. Margo. 2004b. *The economic impact of the 1960s riots: Evidence from property values*, Working paper, vol. 10493. Cambridge, MA: National Bureau of Economic Research.

Curiel, R., R. Cooperman, and G. Arici. 1990. *The Venetian ghetto*. New York: Rizzoli.

Cutler, D., and E. Glaeser. 1997. Are ghettos good or bad? *Quarterly Journal of Economics* 112: 827–872.

Cutler, D., E. Glaeser, and J. Vigdor. 1999. The rise and decline of the American ghetto. *Journal of Political Economy* 107: 455–506.
DeVos, G., and H. Wagatsuma (eds.). 1966. *Japan's invisible race: Caste in culture and personality*. Berkeley: University of California Press.
Drake, S., and H. Clayton. 1945. *Black metropolis: A study of Negro life in a Northern city*. Chicago: University of Chicago Press.
Echenique, F., and R. Fryer. 2004. *On the measurement of segregation*, Working paper. Cambridge, MA: Department of Economics, Harvard University.
Hane, M. 1982. *Peasants, rebels, and outcasts: The underside of modern Japan*. New York: Pantheon.
Hirsch, A. 1983. *Making the second ghetto: Race and housing in Chicago, 1940–1960*. New York: Cambridge University Press.
Johnston, R., J. Forrest, and M. Poulson. 2002. Are there ethnic enclaves/ghettos in English cities? *Urban Studies* 39: 591–618.
Massey, D., and N. Denton. 1993. *American apartheid: Segregation and the making of the underclass*. Cambridge, MA: Harvard University Press.
Osofsky, G. 1971. *Harlem: The making of a ghetto – Negro New York, 1890–1930*. New York: Harper and Row.
Peach, C. 1996. Does Britain have ghettos? *Transactions of the Institute of British Geographers* 21: 216–235.
Wacquant, L. 2004. Ghetto. In *International encyclopedia of social and behavioral sciences*, ed. N. Smelser and P. Baltes. London: Pergamon Press.
Western, J. 1981. *Outcast Cape Town*. Minneapolis: University of Minnesota Press.
Wilson, W. 1987. *The truly disadvantaged: The inner city, the underclass, and public policy*. Chicago: University of Chicago Press.
Wirth, L. 1928. *The ghetto*. Chicago: University of Chicago Press.
Zhou, M. 1992. *Chinatown: The socioeconomic potential of an urban enclave*. Philadelphia: Temple University Press.

Giblin, Lyndhurst Falkiner (1872–1951)

M. Harper

Giblin was born and died in Hobart, Tasmania. Trained in mathematics and statistics at King's College, Cambridge, he became teacher, goldminer, fruit-grower, Labor politician in Tasmania, and soldier before beginning his career as statistician/economist in 1919. In his official positions as Tasmanian Government Statistician (1919–28), Acting Commonwealth Statistician (1931–2), Ritchie Research Professor of Economics at the University of Melbourne (1929–39), member of the Committee of Enquiry into the Australian Tariff (1927–9), of the Commonwealth Grants Commission (1933–6), the Commonwealth Bank Board (1935–42) and chairman of the Finance and Economic Policy Committee (1939–46), he exercised significant influence on economic policy-making in Australia. He shepherded the small band of Australian economists to preserve cohesion within the profession, provided links with governments, and endeavoured to raise public awareness of the nature and dimensions of economic problems.

In the course of his quantitative work, Giblin dealt with the economic choices faced by a politically federated nation, intent on economic development, but with an open economy subject to the vicissitudes of international trade and capital movements. As a basis for Federal financial relations, he pioneered the measurable concepts of relative taxable capacity and severity of taxation. He attempted to measure the 'excess costs' of protection and their redistributive effects. In 1929, while demonstrating the repercussive effects on Australian incomes of adverse movements in the terms of trade, he formulated a first version of the foreign trade multiplier. During the 1940s he devised the money control tool of requiring the trading banks to hold special deposits with the Commonwealth Bank.

Selected Works

1929. (With Brigden, J.B., D.B. Copland, E.C. Dyason , and C.H. Wickens). *The Australian tariff: An economic enquiry.* Melbourne: Melbourne University Press, (Economic Series No. 6).
1930a. State disabilities – with special reference to Tasmania: A memorandum submitted to the Committee of Public Accounts as evidence of Tasmanian disabilities. Appendix J, *The case for Tasmania, 1930.* Hobart: Government Printer.

1930b. *Australia, 1930.* Melbourne: Melbourne University Press, (Economic Series No. 8).
1951. *The growth of a central bank: The development of the Commonwealth Bank of Australia 1924–1945.* Melbourne: Melbourne University Press.

Bibliography

Copland, D., ed. 1960. *Giblin: the scholar and the man*. Melbourne: F.W. Cheshire.

Gibrat, Robert Pierre Louis (1904–1980)

David E. R. Gay

Keywords
Gibrat, R. P.L.; Gibrat's Law; Econometric Society

JEL Classifications
B31

Gibrat was born on 23 March 1904 in Lorient, France, and died on 13 May 1980 in Paris. He studied at Saint Louis de Paris, as well as in Rennes, Lorient and Brest, and in 1922 he entered the Ecole Polytechnique to become a mining engineer. He received a bachelor's degree in science and a doctorate in law from the University of Paris. He was a technical consultant in private firms before being named director of electricity in the Ministry of Public Works, 1940–42. He became Secretary of State for Communications under the Laval government but resigned after the Allied invasion of North Africa. After the Liberation he was chief engineer of mines. He was consulting engineer for French Electric on tidal energy (1945–68) and served as Director General for atomic energy (Indatom, 1955–74), president of the scientific and technical committee of Euratom (1962), and as a consulting engineer for Central Thermique, 1942–80. He taught at Ecole des Mines from 1936 to 1968. He served as President of the French Society of Electricians, Vice President and President of the Civil Engineers of France, President of the Statistical Society of Paris (1966), President of the French Statistical Society, President and Honorary President of the Technical Committee for the Hydrotechnical Society of France, President of the French Meteorological Society (1969), Honorary President of the World Federation of Organizations of Engineers, and President of the French Section of the American Nuclear Society. Gibrat was the author of reports to the Academy of Sciences, some 100 professional articles and two books (on economics and tidal energy), and a Knight of the Legion of Honour.

His major contribution to economics is known as Gibrat's Law. This states that the expected growth rate for a firm is independent of its size. Gibrat's Law has been successfully tested by French and American investigators, among others. His famous economics work, *Les inégalités économiques,* was published in 1931. For his contributions to economics and mathematical economics he was elected a Fellow of the Econometric Society in 1948.

Selected Works

1930. Une loi des répartitions économiques: l'effet proportionnel. *Bulletin de la Statistique générale de la France et du Service d'observations des prix* 19: 469.
1931a. *Les inégalités économiques*. Paris: Sirey.
1931b. Les inégalités économiques. *Revue de l'industrie minérale* 15.
1935. La science économique. Méthodes et philosophie. *Actualités scientifiques et industrielles*. Paris.

Gibrat's Law

José Mata

Abstract

Gibrat's law states that a great deal of the evolution of firm size distribution over time is due to the action of chance. While chance is still seen as an important driver of the size of firms, recent studies call for chance to be supplemented by some more structured models in order to explain the observed patterns of firm size evolution.

Keywords

Concentration; Firm growth; Firm size; Firm size distribution; Gibrat's law; Heteroskedasticity

JEL Classifications

M2

In one of the first studies on firm size distribution (FSD) Gibrat (1931) observed that the size of firms followed the lognormal distribution very closely, from which he concluded that firms' rate of growth ought to be a random process. In particular, he reasoned that growth should not depend on the initial size of firms, as such a process would inevitably produce a lognormal distribution. This assertion became known as Gibrat's law.

If we denote the size of the firm at time t by S_t and the proportional growth between t and $t-1$ by ε_t

$$(S_t - S_{t-1})/S_{t-1} = \varepsilon_t$$

$$\begin{aligned} S_t &= S_{t-1}(1+\varepsilon_1) \\ &= S_0(1+\varepsilon_1)(1+\varepsilon_2)...(1+\varepsilon_t), \end{aligned}$$

taking logs and using the approximation $\log(1+\varepsilon_t) \approx \varepsilon_t$ leads to

$$\log S_t = \log S_0 + \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_t.$$

As $t \to \infty$, $\log S_0$ becomes less important relative to $\log S_t$ and, if ε_t is drawn from a normal distribution with mean μ and variance σ^2, $\log S_t$ can be approximated by a normal distribution with mean μt and variance $\sigma^2 t$. The result that the variance of the FSD is bound to increase over time due to the sole action of chance is probably responsible for the popularity of Gibrat's law in industrial organization, as it provides a nice explanation for the observed empirical regularity that industrial concentration increased over time. Other fields in which the application of Gibrat's law has been discussed include those of distributions of income (Sahota 1978) and the size of cities (Eeckhout 2004).

It is not surprising that soon after the publication of Gibrat's book, different studies tried to test the law empirically (see Sutton 1997, for a survey). Some proceeded to compare the observed firm size distribution with the lognormal distribution, others analysed the relationship between firm size and firm growth. By and large, the results coincided. The firm size distribution seemed to conform well to the lognormal distribution, and firm growth seemed to be largely independent of firm size.

The studies in this first wave typically used data that were readily available in public sources, that is, data on the largest firms in the economy. In an influential study published in the early 1960s, however, Mansfield (1962) collected data on 'practically all firms' in three American industries in different time periods and analysed the relationship between the size and the growth of firms. He suggested that different interpretations regarding the extent to which Gibrat's law was applicable were possible, and tested the validity of the law according to these different interpretations. According to the first interpretation Gibrat's law would hold for all firms including those that exit; with this interpretation, a negative relationship between initial size and growth was discovered in the majority of the samples that were considered. The second interpretation posited that the

law would hold only for those firms that had not exited; a significant relationship showed up only in a minority of samples. The third interpretation stated that the law would hold only for those firms whose size exceeded a given threshold. Setting this threshold as the minimum efficient scale in the industry, Mansfield found no significant relationship in any of the samples. Yet, even with this restricted interpretation, Mansfield's samples failed to pass a second test of Gibrat's law, that the variance of growth would be independent of size.

The topic did not attract much attention during the rest of 1960s and the 1970s. When it again came under scrutiny in the mid- to late 1980s, new data sources had became available. These new data sources covered many more firms than before, providing a much better coverage of the smallest firms in the economy. Furthermore, their longitudinal dimension, which allowed researchers to follow firms over time, led to the discovery that the entry, exit and growth movements that were taking place in most industries of developed countries were of a previously unsuspected large magnitude (see Caves 1998, for a survey).

Concerning Gibrat's law, the major consequence was that attention was this time mostly drawn to the relationship between firm growth and size, and problems of sample selection became routinely addressed using econometric techniques that had meanwhile become available. Whether or not sample selection was taken into account, however, the results of the studies using these recently developed databases suggested that firm growth was not independent of firm size, smaller firms growing faster than their larger counterparts (for example, Evans 1987). Another systematic concern of this literature was heteroskedasticity: most studies found that large firms display a less variable pattern of growth than do smaller units. Although consistent with the idea that the diversification associated with size reduces risk, this is a pattern that does not conform well to Gibrat's postulate.

A negative relationship between size and growth does not imply that concentration does not increase. This point can be clearly seen by imagining that there are firms of only two sizes, small and large, present in equal numbers in an economy. If those firms that are small in one period become large in the next period and vice versa, the overall distribution remains constant despite the obvious relationship between growth and size. There have been few studies in this new wave that have specifically examined the FSD. McCloughan (1995) simulated the effect of different violations of Gibrat's postulates upon the development of market structure and concluded that the nature of the size–growth relationship (in contrast to the effect considering entry and exit) was the most important determinant of the evolution of concentration.

In one of the few studies that have used these new comprehensive data-sets to analyse the actual development of the FSD, Cabral and Mata (2003) found that the FSD is considerably more skewed than the lognormal in the earliest years, but gradually approaches it as firms get older. Convergence to the lognormal is what would be expected from a Gibrat process, and the fact that the lognormal is approached from a more skewed distribution may seem to be unimportant from the strict standpoint of Gibrat's law, as the law posits that the starting point does not matter in the long run. The finding, however, creates an additional challenge: if we are to rely on random forces to explain the evolution of the FSD, what can possibly explain its starting position? How can this be part of a theory of the evolution of firm size, and how can it coexist with models such as Jovanovic's (1982), in which skewedness emerges gradually as firms learn about their abilities? One possibility is that the size of firms at start-up is the minimum of two sizes: a size to be achieved in the long run – determined by the ability of the entrepreneur in the spirit of Lucas (1978) – and a short-run size, given by some constraint. Cabral and Mata suggested that this constraint could to be a financial one, but other constraints might do the job as well.

An unaddressed question is the extent to which the FSD converges to a position that depends on a pre-existing distribution of abilities and to what extent abilities are a product of explicit decisions made by firms as to the

learning process (Ericson and Pakes 1995). Another question pertains to the appropriate level of analysis. Machado and Mata (2000) report that failure to control for industry-specific conditions leads to a significantly greater departure from the lognormal than when these conditions are controlled for. It is also not obvious that the FSD should be governed by the same forces and evolve along the same lines irrespectively of the specific competitive conditions in the industry (Sutton 1997), but little work has been done on how these conditions affect the FSD. Perhaps the streams of the literature that has given more attention to the evolution of industries is that following the work by Klepper and Graddy (1990), which shows that, over their life cycle, industries exhibit significant variation, namely, with respect to the changes in the number of firms and their patterns of growth. The implications of these changes to the evolution of the FSD of industries are still under-explored.

See Also

- ▶ Firm-Level Employment Dynamics
- ▶ Growth and Learning-By-Doing
- ▶ Lognormal Distribution

Bibliography

Cabral, L., and J. Mata. 2003. On the evolution of the firm size distribution: facts and theory. *American Economic Review* 93: 1075–1090.

Caves, R.E. 1998. Industrial organization and new findings on the turnover and mobility of firms. *Journal of Economic Literature* 36: 1947–1982.

Eeckhout, J. 2004. Gibrat's law for (all) cities. *American Economic Review* 94: 1429–1451.

Ericson, R., and A. Pakes. 1995. Markov-perfect industry dynamics: A framework for empirical work. *Review of Economic Studies* 62: 53–82.

Evans, D. 1987. Tests of alternative theories of firm growth. *Journal of Political Economy* 95: 657–674.

Gibrat, R. 1931. *Les Inegalités economiques: applications aux inégalités des richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle: la loi de l'effect proportionnel*. Paris: Sirey.

Jovanovic, B. 1982. Selection and evolution of industry. *Econometrica* 50: 649–670.

Klepper, S., and E. Graddy. 1990. The evolution of new industries and the determinants of market structure. *Rand Journal of Economics* 21: 27–44.

Lucas, R.E. 1978. Size distribution of business firms. *Bell Journal of Economics 9*: 508–523.

Machado, J., and J. Mata. 2000. Box-Cox quantile regression and the distribution of firm sizes. *Journal of Applied Econometrics* 15: 253–274.

Mansfield, E. 1962. Entry, Gibrat's law, innovation, and the growth of firms. *American Economic Review* 52: 1023–1051.

McCloughan, P. 1995. Modified Gibrat growth, entry, exit and concentration development. *Journal of Industrial Economics* 43: 405–433.

Sahota, G.S. 1978. Theories of personal income distribution: A survey. *Journal of Economic Literature* 16: 1–55.

Sutton, J. 1997. Gibrat's legacy. *Journal of Economic Literature* 35: 40–59.

Gide, Charles (1847–1932)

Roger Dehem

Gide was born at Uzés (Gard) and died in Paris. A strong Huguenot family tradition marked his austere and moralizing personality. His doctoral thesis was on *Le droit d'association en matiére religieuse*. He was a founder of the *Revue d'Economie Politique* (1887). He taught at the Faculty of Law in Bordeaux (1874), in Montepellier (1880), in Paris (1898) and at the Collége de France (1919).

Best known for the *History of Economic Doctrines* he wrote with Charles Rist, his *Principes* and his *Cours d'économie politique*, Gide's claim to creativity lies in the field of social ethics. In reaction to Bastiat, whose works introduced him to economics, Charles Gide drew inspiration from Fourier and Robert Owen to become a moral critic of the competitive system. Against the self-seeking competitive spirit, he opposed the values of solidarity and cooperation. Like J.S. Mill, Gide drew a distinction between the realm of natural necessity and that of human volition. The principle of solidarity and cooperation, as a moral duty, should

ultimately supersed the struggle of man against man. As the founder of the Ecole de Nîmes, a Huguenot intellectual centre, Gide remained all his life a preacher of the cooperative gospel. Like Fourier, he was a visionary of a better world to come; like J.S. Mill, he was aware of the historical relativity of human institutions. His artist's mind was unwilling to bend to the canons of any rigid science. The purpose of economics was to enlighten the road and inspire the endeavour to a better world, though his message of hope was tempered by a measure of scepticism. In international relations he was a pacifist.

Selected Works

1872. *Le droit d'association en matiére religieuse*. Paris: Faculté de droit, thesis.

1884. *Principes d'économie politique*. Paris: Larose et Forcel. 26th ed, 1931.

1887. La notion de valeur dans Bastiat au point de vue de la justice distributive. *Revue d'économie politique* 1(3).

1894–9. Proudhon (III, 237–8); J.B. Say (III, 357–8); Saint-Simon (III, 346–7); Solidarity (III, 444–5); J.B. Say (III, 486–7). In *Dictionary of political economy*, ed. R.H. Inglis Palgrave. London: Macmillan.

1904. *Les sociétés coopératives de consommation*. Paris: Colin.

1905. *Les institutions du progrés social au début du XXe siècle*. Paris: Larose et Tenin.

1907. Economic literature in France at the beginning of the twentieth century. *Economic Journal* 17: 192–212.

1915. (With Ch. Rist.) *A history of economic doctrines*. London: G. Harrap. 2nd ed, 1948.

1920. *Des institutions en vue de la transformation ou de l'abolition du salariat*. Paris: Giard.

1924. *Fourier, précurseur de la coopération*. Paris: Association pour l'Enseignement de la Coopération.

1928. *L'école de Nîmes*. Paris: Association pour l'Enseignement de la Coopération.

1930. *Les colonies communistes et coopératives*. Paris: Association pour l'Enseignement de la Coopération.

1930–35. Boyve (II, 670); Cooperation (IV, 376–81). In *Encyclopaedia of the social sciences*. London: Macmillan & Co.

1931. (With W. Oualid.) Le bilan de la guerre pour la France. In *Histoire économique et sociale de la guerre mondiale*. Paris: Presses Universitaires de France.

1932. *La solidarité*. Paris: Presses Universitaires de France.

Giffen, Robert (1837–1910)

Murray Milgate

Keywords

Giffen, R.; Measurement; Royal Economic Society; Royal Statistical Society; Statistical comparisons

JEL Classifications

B31

Robert Giffen's name seems likely to be known by students of economics for generations to come in relation to the famous result in the theory of consumer demand which bears his name but about which, so far as can be determined, he had nothing to say. Marshall originated the tradition when he associated the result with Giffen's name in the third edition of his *Principles* in 1895 (p. 208).

Giffen was born in Lanarkshire. At the age of 13 he was apprenticed to a solicitor in Strathaven, and continued in the same vocation until 1860 (though during the last seven years of this period he resided and worked in Glasgow). Still only 23 years old, Giffen struck out on a career in journalism – in which he was to be successful in

establishing his reputation in economic circles of the day. He begun as a sub-editor for the *Stirling Journal*, moved to London in 1862 to work at the *Globe*, transferred to the *Fortnightly Review* in 1866, and in 1868 became assistant editor at *The Economist* – a post at which he remained until his next change of vocation in 1876. He was also city editor at the *Daily News* between 1873 and 1876. Giffen's third and final career was as a professional civil servant, first as chief of the statistical department at the Board of Trade, and then in 1882 as its Assistant Secretary. He retired from the civil service at the age of 60. Giffen served on numerous royal commissions (including the Gold and Silver Commission of 1886–8); he was editor of the *Journal of the Royal Statistical Society* (1876–91), President of that society (1882–4), twice presided over the economics section of the British Association (1887 and 1901), and was one of the founders of the Royal Economic Society. In short, he was one of those figures encountered frequently in British economics whose not inconsiderable power and prestige appears to be disproportionate to their actual contribution to economic science.

In so far as he was primarily a statistician, Giffen's work did attempt to alert economists to the dangers of theory without measurement. His presidential address to the Royal Statistical Society in 1882 was devoted to the subject, and in 1901 as president of Section F of the British Association (his second term in that office) he returned to the same theme (see 1904, vol. 2, chs 13 and 28). Indeed, according to Higgs in his edition of this *Dictionary* (1925), Giffen's statistical prowess was one of the factors which helped to secure the respect of theorists. His article on international statistical comparisons in the *Economic Journal* (1892b), for example, can be singled out for special mention since it treats for the first time a problem which has still not been adequately resolved. Of course, it was not always the case that Giffen's careful mustering of the statistical evidence allowed him, any more than the theorists, to avoid the pitfalls of making predictions which subsequent experience has proven to be silly – witness his claim that the whole protectionist school would die out within a decade (1898, p. 16).

However, in the final analysis it is in Giffen's attempts to provide reasonably accurate measurements of indicators like wage rates, economic growth (see 1884), and national product (1889) that one should isolate his main contribution. While it is true that subsequent work in this field has advanced well beyond Giffen's early efforts, he remains one of the pioneers of applied economics in its modern sense.

It seems that Giffen was also a strong supporter of a Channel tunnel: not for one between England and France, but between Ireland and England. He died on 12 April 1910 and is buried in Strathaven.

Selected Works

1872. (With B. Cracroft.) *American railways as investment.* London.
1873. The production and movement of gold since 1848. In Giffen (1880).
1877. *Stock exchange securities.* London: G. Bell & Sons.
1880. *Essays in finance.* First Series, vol. 1. London: G. Bell & Sons.
1884. *The progress of the working class in the last half century.* London: G. Bell & Sons.
1886. *Essays in finance*. Second Series, vol. 2. London: G. Bell & Sons.
1887. The recent rate of material progress in England. Address as President of Section F of the British Association. *Journal of the Royal Statistical* Society 50, 615–647. Reprinted in Giffen (1904), vol. 2.
1889. *The growth of gapital*. London: G. Bell & Sons.
1892a. *The case against bimetallism*. London: G. Bell & Sons.
1892b. On international statistical comparisons. *Economic Journal* 2(June), 209–238.
1898. Protection for manufactures in new countries. *Economic Journal* 8(March), 3–16.
1904. *Economic inquiries and studies*, 2 vols. London: G. Bell & Sons.

Giffen's Paradox

John Nachbar

Keywords

Comparative statics; General equilibrium; Giffen good; Giffen preferences; Giffen, R; Giffen's paradox; Income effect; Law of demand; Preference externalities; Revealed preference; Slutsky equation; Substitution effect

JEL Classifications

D12

Giffen's paradox refers to the possibility that standard competitive demand, with nominal wealth held constant, can be upward sloping, violating the law of demand. From the Slutsky equation, Giffen's paradox arises if and only if a good is inferior and the income effect is larger than the absolute value of the substitution effect. A *Giffen good* is a good for which Giffen's paradox can arise. *Giffen preferences* are preferences that can exhibit Giffen's paradox. For explicit examples of Giffen preferences, see Moffatt (2002) and Sorensen (2005).

The term 'Giffen's paradox' originates in a passage in Marshall (1920), which credits the statistician Robert Giffen (1837–1910) with observing a failure of the law of demand in the market for bread. The widespread association of Giffen's paradox with potatoes during the Irish potato famine of the 1840s may have originated in Samuelson (1964). A number of authors have since argued that potatoes were not, in fact, a Giffen good for Irish potato farmers (for example, Dwyer and Lindsay 1984; Rosen 1999). For more on the tortured intellectual history of Giffen's paradox, see Walker (1987). For empirical evidence that Giffen goods may exist, see Baruch and Kannai (2001) and Jensen and Miller (2002).

In thinking about Giffen's paradox, bear in mind four points. First, the budget constraint forces a crude form of compliance with the law of demand: as the price of a good goes to infinity, consumption of that good must go to zero. Conversely, under standard assumptions on preferences (such as monotonicity), as the price of a good goes to zero, demand for the good becomes large. Thus, under standard assumptions, the graph of demand for a Giffen good, with price on the vertical axis, is roughly Z–shaped.

Second, whether Giffen's paradox arises for aggregate demand, summed across consumers, depends on the wealth distribution as well as on individual preferences. This point is a variation on the preceding one. For any given consumer, if prices are held fixed, consumption of any good must fall for large enough drops in nominal wealth, which implies that no good is inferior for all wealth levels. Therefore, even if consumers have the same Giffen preferences, if consumers have different wealth levels, then, over any given price interval, the law of demand may hold for some consumers even if it is violated for others. As a consequence, even if all consumers have the same Giffen preferences, aggregate demand may obey the law of demand. A striking example is due to Hildenbrand (1983): if there is a continuum of consumers all of whom have the same preferences and if nominal wealth is uniformly distributed on an interval containing zero, then aggregate demand cannot exhibit Giffen's paradox.

Third, in a general equilibrium (GE) model, with endogenous wealth, the comparative statics of Giffen goods can run counter to standard textbook intuition. Consider an exchange economy (no production) with two goods and only one consumer. In the equilibrium of this trivial economy, the consumer eats her own endowment. Equilibrium relative prices (which are well defined even though there is no trade) are given by the slope of the consumer's indifference curve through her consumption/endowment point. Naive textbook supply and demand analysis predicts that, if demand for good 1 is upward sloping, then an increase in the endowment of good 1 results in a higher equilibrium price (if we assume that the price of good 2 is constant). One

can easily show, however, that in this economy an increase in the endowment of good 1 can increase its equilibrium price only if the good is *normal,* hence not Giffen. In fact, if the good is Giffen then the endowment increase causes the price to fall by so much that nominal wealth falls (causing the demand curve to shift *out,* since the good is inferior), even though the endowment increase makes the consumer better off.

The critical feature of this example is that individual demand automatically obeys the weak axiom of revealed preference, and in exchange economies the weak axiom implies the law of demand for *endogenous* wealth demand near equilibrium. Thus, if preferences are Giffen, endogenous wealth demand slopes down at equilibrium even though fixed wealth demand slopes up, and it is endogenous wealth demand that drives GE comparative statics. This analysis generalizes to multi-consumer economies, with active trade, provided aggregate, endogenous wealth demand satisfies the weak axiom (see Nachbar 2002); the basic intuition goes back to Hicks (1939). Note, however, that in multi-consumer economies aggregate demand does not automatically satisfy the weak axiom.

For other work that investigates the behaviour of Giffen goods in environments richer than those usually considered in textbook treatments, see Barzel and Suen (1992) and Rosen (1999).

Finally, in standard economics usage, Giffen's paradox refers both to a phenomenon – failure of the law of demand for standard competitive demand – and to a particular mechanism underpinning that phenomenon – income effects. Giffen's paradox is, however, sometimes conflated with similar phenomena arising from quite different mechanisms, often based on preference externalities of some form. A classic citation is Leibenstein (1950). For an interesting application, see Pesendorfer (1995).

See Also

► Comparative Statics
► Law of Demand

Bibliography

Baruch, S., and Y. Kannai. 2001. Inferior goods, Giffen goods, and shochu. In *Economics essays: A Festschrift for Werner Hildenbrand*, ed. G. Debreu, W. Neuefeind, and W. Trockel. Heidelberg: Springer.

Barzel, Y., and W. Suen. 1992. The demand curves for Giffen goods are downward sloping. *The Economic Journal* 102: 896–905.

Dwyer, G. Jr., and C. Lindsay. 1984. Robert Giffen and the Irish potato. *American Economic Review* 74: 188–192.

Hicks, J. 1939. *Value and capital*. Oxford: Clarendon Press.

Hildenbrand, W. 1983. On the law of demand. *Econometrica* 51: 997–1018.

Jensen, R., and N. Miller. 2002. *Giffen behaviour: Theory and evidence*. Working Paper No. RWP02-014. Kennedy School of Government, Harvard University.

Leibenstein, H. 1950. Bandwagon, snob, and Veblen effects in the theory of consumers' demand. *Quarterly Journal of Economics* 64: 283–207.

Marshall, A. 1920. *Principles of economics,* 8th ed. London: Macmillan.

Moffatt, P. 2002. Is Giffen behaviour compatible with the axioms of consumer theory? *Journal of Mathematical Economics* 27: 259–267.

Nachbar, J. 2002. General equilibrium comparative statics. *Econometrica* 70: 2065–2074.

Pesendorfer, W. 1995. Design innovation and fashion cycles. *American Economic Review* 85: 771–792.

Rosen, S. 1999. Potato paradoxes. *Journal of Political Economy* 107: 294–313.

Samuelson, P. 1964. *Economics,* 5th ed. New York: McGraw-Hill.

Sorensen, P. 2005. *Simple utility functions with Giffen demand*. Working paper. Department of Economics, University of Copenhagen.

Walker, D. 1987. Giffen's paradox. In *The new Palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman, vol. 2. Basingstoke: Palgrave.

Gifts

C. A. Gregory

A gift, according to the *Concise Oxford Dictionary*, is a 'voluntary transference of property; thing given, present, donation'. For most economists, especially those familiar only with industrial capitalist economies, this is all that need be said on the matter: it is obvious what gift exchange is and there is nothing to be explained.

The only problem the phenomenon of exchange poses for the economist is that of 'value' and this arises in the context of commodity exchange.

For the anthropologist, however, the phenomenon of exchange poses questions about the nature of gift exchange. These lie at the centre of the discipline and the topic has been the subject of much theoretical debate. Anthropologists stress that while gifts appear to be voluntary, disinterested and spontaneous, they are in fact obligatory and interested. It is this underlying obligation that anthropologists seek to understand: What is the principle whereby the gift received has to be repaid? What is there in the thing given that compels the recipient to make a return?

It is clear that the one economic category – exchange – means fundamentally different things to different people and that these contrary perceptions of the exchange process have given rise to quite distinct theoretical traditions. The reasons for this are to be found in the historical conditions which gave rise to the development of the academic disciplines of economics and anthropology. The history of economic thought must be understood with reference to the development of mercantile and industrial capitalism in Europe; the development of anthropological theorizing, on the other hand, must be situated in the context of the imperialist expansion of European capitalism and especially the colonial conquest of Africa and the Pacific towards the end of the 19th century. The fact that economists have been preoccupied with commodity exchange whilst anthropologists have been primarily concerned with gift exchange simply reflects the fact that the modern European economy is organized along very different lines from the indigenous economies of Africa, the Pacific and elsewhere. The data anthropologists have collected from these countries over the past one hundred years has revolutionized our understanding of tribal economy and the theory of the gift; their theoretical reflections on this data constitutes a major contribution to the theory of comparative economic systems and also to the theory of development and underdevelopment. This anthropological literature takes us far beyond the superficial dictionary definition of the gift and raises important questions about the seemingly unrelated issue of shell money; interestingly it also brings us back to the original meaning of the word as 'payment for a wife' and 'wedding' found in *The Oxford Dictionary of English Etymology*.

Anthropological accounts of gift giving first began appearing towards the end of the 19th century; by the end of World War I a large quantity of data had been collected. The most spectacular accounts came from the Kwakiutl Indians of the northwest coast of America and from the Melanesian Islanders of the Milne Bay District of Papua New Guinea. Among the Kwakiutl vast amounts of valuable property (mainly blankets) are ceremonially destroyed in a system called 'potlatch' (Boas 1897). In the potlatch system the prestige of an individual is closely bound up with giving: a would-be 'big-man' or 'chief' is constrained to give away or to destroy everything he possesses. The principles of rivalry and antagonism are basic to the system and people compete with one another, each trying to outgive the other in order to gain prestige. The status and rank of individuals and clans is determined by this war of property. In Papua New Guinea, the classic home of competitive gift exchange, the instruments of 'gift warfare' are food (Young 1971) and shells of various shapes and sizes (Leach and Leach 1983). These are not destroyed but transacted by status seekers according to complicated sets of rules which we are only now beginning to understand. Most Papua New Guinean societies are without any form of ascribed status and the egalitarian ideology of these societies means that competitive gift giving is primarily concerned with the maintenance of equal status rather than dominance. Staying equal is, as Forge (1972) has pointed out, an extremely onerous task requiring continual vigilance and effort: perfect balance is impossible to achieve as the temporal dimension of gift exchange necessarily introduces status inequalities. Perhaps the most complicated gift exchange system in Melanesia is the Rossel Island 'monetary system'. This was first described by Armstrong (1924) and has recently been restudied by Liep (1983). On Rossel Island there are two kinds of 'shell money' *ndap* and *ko*. A single unit of *ndap* is a polished piece of *spondylus* shell a few millimetres thick, having an area varying from

2 to 20 square centimetres and roughly triangular in shape. A single unit of *ko* consists of ten pieces of *chama* shell of roughly the same size and thickness with a small hole in the centre for binding them together. Each shell group contains some forty-odd hierarchical divisions. What is unusual about these divisions is that they have rank rather than value, that is, they are ordinally related rather than cardinally related. For example, the relationship of a big *ndap* shell to a small one is analogous to that between an ace of hearts and a two of hearts rather than that between a dollar and a cent.

The publication of Armstrong's (1928) ethnography of Rossel Island and Malinowski's (1922) now classic description of an inter-island gift exchange system called *kula* sparked off a debate about the nature of 'shell money' which still rages today. This debate is kept alive not from an antiquarian interest in 'archaic' money systems but because these gift exchange systems are still flourishing despite their incorporation into the world capitalist economy (Macintyre and Young 1982; Gregory 1980, 1982). On Rossel Island, for example, not only are the *ndap* and *ko* shells still transacted as gifts according to the complicated rules of old, but the demand for Rossel Island *chama* shells for use in the flourishing *kula* gift exchange system of neighbouring islands has transformed Rossel Island into a major commodity producer and exporter of *chama* shells (Liep 1981; 1983).

These facts raise conceptual questions about the difference between gift exchange and commodity exchange, and theoretical and empirical questions about the nature of the interaction between them. Neoclassical economics answers these questions within a framework that employs the universalist and subjectivist concept 'goods', a category which, by definition, cannot explain the particularist and objective nature of gift and commodity exchange (Gregory 1982). A 'gift' therefore becomes a 'traditional good' and highly questionable psychological criteria are used to distinguish this from a 'modern good'. For example, Einzig argues that 'the intellectual standard' of people in tribal societies 'is inferior and their mentality totally different from ours' (1948, p. 16); Stent and Webb (1975, p. 524) argue that 'traditional' consumers in Papua New Guinea are on the bliss point of their indifference curves. A further difficulty economists have with the problem of contrasting economic systems – and this is not restricted to neoclassical economic thought – is the habit of beginning an argument with an analysis of barter in an 'early and rude state of society'. The barter economies of these theories are figments of a Eurocentric imagination that bear no resemblance at all to actual tribal economies. Economic anthropologists have been making this point for over fifty years but without much success (Malinowski 1922, pp. 60–61; Polanyi 1944, pp. 44–5). What is needed, then, is an empirically based theory of comparative economy. The foundations of such a theory were laid by Marx (1867) but the rise to dominance of neoclassical theory precluded any further development of the theory of comparative economy within the economics discipline. The theoretical advances have come from without and have been made by anthropologists, sociologists and economic historians.

The outstanding contribution to the 20th-century literature is undoubtedly Mauss's *The Gift: Forms and Functions of Exchange in Archaic Societies*, first published in French in 1925 as 'Essai sur ie don, forme archaique de l'echange' in Durkheim's journal, *L'Année Sociologique*. Mauss (1872–1950) was Durkheim's nephew and became a leading figure in French sociology after his uncle's death. His essay on the gift is a remarkable piece of scholarship. Not only did he survey all extant ethnographic data on gift giving from Melanesia, Polynesia, northwest America and elsewhere, he also examined the early literature from Ancient Rome, the Hindu classical period and the Germanic societies. His essays conclude with a critique of western capitalist society by drawing out the moral, political, economic and ethical implications of his analysis.

The key to understanding gift giving is apprehension of the fact that things in tribal economies are produced by non-alienated labour. This creates a special bond between a producer and his/her product, a bond that is broken in a capitalist society based on alienated wage-labour. Mauss's analysis focused on the 'indissoluble bond' between things and persons in gift economies and argued that 'to

give something is to give a part of oneself' (1925, p. 10). Gifts therefore become embodied with the 'spirit' of the giver and this 'force' in the thing given compels the recipient to make a return. This does not exist in our system of property and exchange which is based on a sharp distinction between things and persons, that is, alienation (1925, p. 56). The wage-labourer in a capitalist society gives a 'gift' which is not returned (1925, p. 75). Capitalism for Mauss then was a system of non-reciprocal gift exchange; a system where the recipients of a gift were under no obligation to make a return gift.

This analysis of the wage-labour contract under capitalism has a Marxian ring about it. However, Mauss was no revolutionary and he drew very different policy conclusions from his analysis of the wage-labour relation. He argued for a welfare capitalism where the state, through its social legislation, provided recompense to the workers for their gifts.

A feature of Mauss's work, and indeed a feature of much early theorizing about the gift, was the evolutionary framework within which the ethnographic data was analysed. The tribal economies studied by anthropologists were seen as living fossils from European pre-history, hence the use of terms such as 'archaic' and 'primitive'. These early theorists, then, were only concerned with the intellectual contribution this data could make to the study of comparative economy. To the extent that they were concerned with the welfare of living people it was the welfare of their European countrymen and women; they were not concerned with policies for the development of tribal peoples.

The other outstanding theorist in this evolutionary tradition was another Frenchman, Claude Lévi-Strauss. His theory of the gift is contained in his *The Elementary Structures of Kinship* (1949). Like Mauss's *The Gift*, Lévi-Strauss's book is an encyclopaedic survey of the ethnographic literature. Its central focus is marriage. In line with a long tradition in anthropology he conceptualizes this as an exchange of women. However, Lévi-Strauss's innovation is to argue that women are the 'supreme gift' and that the incest taboo is the key to understanding gift exchange. The virtual universal prohibition on marriage between close kin, he argues, is the basis of the obligation to give, the obligation to receive, and the obligation to repay.

Lévi-Strauss's theory is an analytical synthesis of literally thousands of ethnographic accounts from the Australian Aborigines, the Pacific and Asia. The original or most elementary form of gift exchange, according to Lévi-Strauss, is 'restricted' exchange where the moieties of a population exchanged sisters at marriage; the second form is 'delayed' exchange where a woman is given this generation and her daughter returned the next; the most advanced form is 'generalized' exchange where one clan gives women to another clan but never receives any in return, the closure of the system being brought about by a circle of giving. In the movement from one stage to another, extra spheres of gift exchange are developed as symbolic substitutes for women. These are needed to maintain the ever widening marriage alliances brought about by the shift from restricted to generalized exchange. This movement from marriage to exchange is an aspect of an opposing movement from exchange to marriage. Lévi-Strauss sees a continuous transition from war to exchange, and from exchange to intermarriage as effecting a transition from hostility to alliance, and from fear to friendship.

Lévi-Strauss's theory has attracted considerable critical attention and has been described by his principal opponent as 'in large measure fallacious' (Leach 1970, p. 111). Whatever its shortcomings his theory nevertheless manages to establish the important link between gift giving and the social organization of kinship and marriage. In other words, he has established a relationship between the obligation to give and receive gifts and the biological and social basis of human reproduction.

While Lévi-Strauss was developing his theory of the gift, an economic historian, Karl Polanyi, was approaching the problem from an altogether different perspective in his classic study, *The Great Transformation* (1944). His problem was the analysis of the emergence of the 'self-regulating market' and in order to grasp the 'extraordinary assumptions' underlying such a system he

developed a theory of comparative economy based on ethnographic and historical evidence.

Polanyi correctly identified the Smithian 'paradigm of the bartering savage', which is accepted as axiomatic by many social scientists, as a barrier to an adequate understanding of non-market economy. In a tribal economy, notes Polanyi, the propensity to truck, barter and exchange does not appear: there is no principle of labouring for remuneration, the idea of profit is banned and giving freely is acclaimed a virtue. How, then, is production and distribution ensured, he asks. Polanyi devoted only ten pages of his book to answering this question but his insights have had a significant impact on anthropologicalthought (see e.g. Dalton and Kocke 1983). Tribal economy, he argued, is organized in the main by two principles; *reciprocity* and *redistribution*. Reciprocity works mainly in regard to the sexual organizations of society, that is, family and kinship, and it is that broad principle which helps to safeguard both production and family sustenance. Redistribution refers to the process whereby a substantial part of all the produce of the society is delivered to the chief who keeps it in storage. This is redistributed at communal feasts and dances when the villagers entertain one another as well as neighbours from other districts.

Reciprocity and redistribution are able to work because of the institutional patterns of *symmetry* and *centricity*. Tribes, says Polanyi, are subdivided along a symmetrical pattern and this duality of social organization forms the 'pendant' on which the system of reciprocity rests. (Lévi-Strauss's restricted exchange model of gift exchange also presupposes dual social organization.) The institution of territorial centricity forms the basis of redistribution.

To these two principles, Polanyi adds a third – *householding*, production for use with *autarky* as its basis – and argues that all economic systems known to us up to the end of feudalism were organized on either the principle of reciprocity, or redistribution, or householding, or some combination of the three. These made use of the patterns of symmetry, centricity and autarky, with custom, law, magic and religion cooperating to induce the individual to comply with the rules of behaviour.

Capitalism, in Polanyi's view, implies the wholesale destruction of these principles and the establishment of free markets in land, money and labour run according to the profit principle. Like Marx, Polanyi sees the emergence of free wage-labour as a commodity as the crucial defining characteristic of capitalism. Labour was the last of the markets to be organized in England and both Marx and Polanyi saw the enclosure movements, especially those at the time of the industrial revolution, as central to this process. Polanyi is more precise in his historiography however. He sees the Poor Law Reform of 1834, which did away with the final obstruction to the functioning of a free labour market, as the beginning of the era of the self-regulating market.

Postwar developments in the theory of the gift have built on the foundations laid by Mauss, Lévi-Strauss and Polanyi. The influential contributions of Godelier (1966, 1973), Meillassoux (1960, 1975) and Sahlins (1972) in particular are heavily indebted to these theorists whose ideas they attempt to develop in the light of Marx's theory of comparative economy. Recent empirical research (e.g. Strathern 1971; Young 1971; Leach and Leach 1983) has provided, and will continue to provide, the basis for new comparative insights into the theory of the gift (Forge 1972).

An important postwar development in the theory of the gift has been the analysis of the impact of colonization and capitalist imperialism on tribal societies.

For the early contributors to this literature the problem was how to explain the process of destruction brought about by capitalism. Paul Bohannan (1959), an American anthropologist with fieldwork experience in West Africa, developed a theory of the impact of money on a tribal economy based on Polanyi's ideas. Commodity exchange, according to Polanyi, is a 'uni-centric economy' because of the nature of 'general purpose money' which reduces all commodities to a common scale. In a tribal society, by way of contrast, the economy is 'multi-centric': there are multiple spheres of exchange, each with 'special purpose' money that

could only circulate within that sphere. Among the Tiv of West Africa, for example, there were three spheres of exchange. The first sphere contained locally produced foodstuffs, tools and raw materials; the second sphere contained non-market 'prestige' goods such as slaves, cattle, horses, prestige cloth *(tuguda)* and brass rods; the third sphere contained the 'supreme gift', women. Bohannan's argument was that the general purpose money introduced by the colonial powers reduced all the various spheres to a single sphere thereby destroying them.

Bohannan's theory was applied to the analysis of the impact of colonization in other parts of the world, Papua New Guinea among others (e.g. Meggitt 1971). While Bohannan's theory makes an important conceptual advance in comparative economy it is now recognized that his theory of the impact of colonization has a number of shortcomings as a description of what happened in West Africa (see Dorward 1976); furthermore, it does not pose the problem to be explained. Today, it is now realized, the problem is not, 'How was the tribal gift economy destroyed?' but rather, 'Why has it flourished under the impact of colonization?'

Take the famous potlatch system, for example. The establishment of a canning industry in the area in 1882 led to a rapid increase in the per capita income of the Kwakiutl, a rapid increase in the number of blankets that could be purchased, and hence a rapid increase in the number of blankets given away in potlatch ceremonies. Before the canning industry was established the largest potlatch consisted of 320 blankets, but during the period 1930–1949 potlatch ceremonies involving as many as 33,000 blankets were recorded (Codere 1950, p. 94). This rapid growth in potlatch occurred despite the institution in 1885 of a law prohibiting the ceremonies. The system has not retained its pristine form, however. Legal and other influences have brought about a variety of outward changes in form but the original purpose of the system still persists: the presentation of a claim to a specific social status (Drucker and Heizer 1967. pp. 47–52).

In Papua New Guinea, to take another example, the establishment of one of the world's largest copper mines in Bougainville has stimulated a flourishing import of shells into the island. The shells are manufactured by the Langalanga people of western Malaita in the Solomon Islands some 1550 kilometres away. The mine has given the people of Bougainville income earning opportunities unavailable to other islanders and they are able to outbid other purchasers for the Langalangan shells. The Langalangans, for their part, have oriented all their production away from local purchasers to the Bougainville market. In Bougainville the shells are used mostly by the Siwai people who give them as marriage gifts and traditional gift exchanges involving land and pigs; they are also used as ornaments (see Connell 1977).

This symbiosis between commercialization and gift exchange is found elsewhere in Papua New Guinea. The famous *kula* gift exchange system in the Milne Bay District still persists despite more than one hundred years of colonization (Leach and Leach 1983). Milne Bay is now something of an economic backwater, its heyday of commercial development being the gold mining era early in this century. Labour is probably one of the area's most important exports today. These migrants maintain close contact with their villages and often send home money, some of which is channelled into *kula* transactions. The migrants, who are senior public servants, entrepreneurs, and politicians, also take their culture with them to the urban areas. The result is that the *kula* ring now extends to Port Moresby, where Mercedes cars and telephones have replaced outrigger canoes and conch shell horns as the principal means of communication.

There is some empirical evidence that appears to contradict the theses that gift exchange has effloresced under the impact of colonization. Prior to the European colonization of West Africa and India these countries were part of a flourishing international cowrie-shell economy. The shells *(cyprae moneta)* were produced in the Maldive Islands of the Indian Ocean and were shipped to West Africa and India where they were used primarily as instruments of exchange

but also for religious and ornamental purposes (Heimann 1980). The cowrie shells were an important and profitable item of international trade in the mercantile era. They were purchased very cheaply in the Maldives – where they grow in great profusion – and exported to India or Europe. The merchants of Europe re-exported them to West Africa where they used them to purchase slaves.

This international shell economy, which had persisted for many centuries, began to collapse around the middle of the 18th century. The supply of shells began to increase rapidly and their price began to fall. For example, in 1865, 1636 tons of cowries were imported into Lagos; by 1878 imports totalled 4472 tons, which was the peak; ten years later imports had fallen to a mere ten tons. Cowrie shell prices (measured in pounds sterling) collapsed over this period. In 1851 two thousand cowries cost 4 s. 9d. but by 1876–79 the price had fallen to 1 s. 0d. (Hopkins 1966; Johnson 1970). By the beginning of the 20th century cowrie shells were no longer current; their place had been taken by the fiat money of the respective colonial government.

This evidence of the destructive impact of colonization only appears to contradict the 'efflorescence of gift exchange' thesis however. The reality is otherwise and the evidence demonstrates the point that exchange is a social relationship which varies depending upon the political and historical context. Objects, such as shells, have many uses, and the historical fact that they have been used as instruments of gift exchange here, as objects of commodity exchange there, and as currency in other places has caused great confusion in the literature. The issue is further confused by the fact that in contemporary Papua New Guinea for example, a shell may be used in all three roles during the same day. The issue can be clarified somewhat by inquiring into the primary role of an exchange object and situating this historically and comparatively in terms of the mode of reproduction of a society. The uniqueness of a place such as Papua New Guinea becomes apparent from this perspective. Papua New Guinea, unlike West Africa or India, was not part of an international mercantile economy prior to European colonization, and as a result commercial exchange transactions were a subordinate and insignificant part of total exchange. Pre-colonial India and West Africa, on the other hand, were highly commercialized: land and labour were freely transacted as commodities with gold and silver commodity monies being used as the principal instruments of exchange. The colonization of West Africa transformed it from being a stateless commodity economy to a state controlled one. This involved a suppression of the stateless commodity monies and their substitution by state fiat money. In India a similar process occurred as the British Government established strong centralized administrative control over numerous weak, corrupt princely states. The destruction of the cowrie shell economy must be seen as part of this process of transition from stateless commodity money to state fiat money. Cowries were the small change of gold and silver. The relationship of cowries to gold and silver, then, finds its counterpart in the relationship of pennies to shillings and pounds. However, whereas the relationship between gold and cowries is determined by production conditions and changes from day to day, the relationship between pounds and pennies is set by government decree and never changes. Where a stable government exists, and the value of money remains constant, it is obvious that a merchant or consumer will prefer to use the latter.

The shells used in West Africa and India, then, were used primarily as instruments of commodity exchange and the term 'shell money' is correct in this context. However, the shells used in the exchange systems of Melanesia and elsewhere were not used as the small change of commodity monies in pre-colonial times. They were used primarily as instruments of gift-exchange and the term 'shell gifts' is more appropriate in this context. Colonization has resulted in the efflorescence of gift-exchange in Melanesia because the colonial state brought an end to tribal warfare and facilitated a transition from fighting with

weapons to fighting with gifts. These gifts take the form of women, shells, food and even money nowadays. These gifts do not involve a 'voluntary transference of property' as the *Oxford English Dictionary* would have it. They are the results of obligations imposed on people struggling to achieve status and wealth in a situation where indigenous systems of land tenure, kinship and marriage are being incorporated into an international economic and political order, over which tribespeople and peasants have little control.

See Also

- ▶ Economic Anthropology
- ▶ Exchange

Bibliography

Armstrong, W.E. 1924. Rossel Island money: A unique monetary system. *Economic Journal* 34: 423–429.

Armstrong, W.E. 1928. *Rossel Island: An ethnological study*. Cambridge: Cambridge University Press.

Boas, F. 1897. In *Kwakiutl ethnography*, ed. H. Codere. Chicago: University of Chicago Press, 1966.

Bohannan, P. 1959. The impact of money on an African subsistence economy. *Journal of Economic History* 19(4): 491–503.

Codere, H. 1950. *Fighting with property*. New York: Augustin.

Connell, J. 1977. The Bougainville connection: Changes in the economic context of shell money production in Malaita. *Oceania* 48(2): 81–101.

Dalton, G., and J. Köcke. 1983. The work of the Polanyi group: Past, present and future. In ed. S. Ortiz. 1983.

Dorward, D.C. 1976. Precolonial Tiv trade and cloth currency. *The International Journal of African Historical Studies* 9(4): 576–591.

Drucker, P., and R.F. Heizer. 1967. *To make my name good: A reexamination of the Southern Kwakiutl Potlatch*. Los Angeles: UCLA Press.

Einzig, P. 1948. *Primitive money*. London: Eyre/Spottiswoode.

Forge, A. 1972. The golden fleece. *Man* 7(4): 527–540.

Godelier, M. 1966. *Rationality and irrationality in economics*. London: New Left Books, 1972.

Godelier, M. 1973. *Perspectives in Marxist anthropology*. Cambridge: Cambridge University Press, 1977.

Gregory, C.A. 1980. Gifts to men and gifts to god: Gift exchange and capital accumulation in contemporary Papua. *Man* 15(4): 626–652.

Gregory, C.A. 1982. *Gifts and commodities*. London: Academic Press.

Heimann, J. 1980. Small change and ballast: Cowry trade and usage as an example of Indian Ocean economic history. *South Asia* 3(1): 48–69.

Hopkins, A.G. 1966. The currency revolution in south-west Nigeria in the late nineteenth century. *Journal of the Historical Society of Nigeria* 3(3): 471–483.

Johnson, M. 1970. The cowrie currencies of West Africa. *Journal of African History* 11(1): 17–49; 11(3), 331–53.

Leach, E.R. 1970. *Lévi-Strauss*. London: Fontana.

Leach, J.W., and E. Leach (eds.). 1983. *The Kula*. Cambridge: Cambridge University Press.

Lévi-Strauss, C. 1949. *The elementary structures of - Kinship*. Trans. London: Eyre and Spottiswoode, 1969.

Liep, J. 1981. The workshop of the Kula: Production and trade of shell necklaces in the Louisade Archipelago. *Folk og Kultur* 23: 297–309.

Liep, J. 1983. Ranked exchange in Yela (Rossel Island). In *The Kula*, ed. J.W. Leach and E. Leach. Cambridge: Cambridge University Press.

MacIntyre, M., and M. Young. 1982. The persistence of traditional trade and ceremonial exchange in the Massim. In *Melanesia: Beyond diversity*, ed. R.J. May and Hank Nelson. Canberra: Australian National University.

Malinowski, B. 1922. *Argonauts of the Western Pacific*. New York: E.P. Dutton, 1961.

Marx, K. 1867. *Capital. Vol. 1: A critical analysis of capitalist production*. Moscow: Progress Publishers, n.d.

Mauss, M. 1925. *The gift*. London: Routledge/Kegan Paul, 1974.

Meggitt, M.J. 1971. From tribesman to peasants: The case of the Mae-Enga of New Guinea. In *Anthropology in oceania*, ed. L.R. Hiatt and C.J. Jayawardena. Sydney: Angus/Robertson.

Meillassoux, C. 1960. Essai d'interprétation du phénoméne économique dans les sociétés traditionelles d'auto-subsistance. *Cahiers d'Etudes Africaines* 4: 38–67.

Meillassoux, C. 1975. *Maidens, meal and money*. Cambridge: Cambridge University Press, 1981.

Ortiz, S. (ed.). 1983. *Economic anthropology: Topics and theories*. New York: University Press of America.

Polanyi, K. 1944. *The great transformation*. New York: Rinehart.

Sahlins, M. 1972. *Stone age economics*. Chicago: Aldine.

Stent, W.R., and L.R. Webb. 1975. Subsistence, affluence and market economy in Papua New Guinea. *Economic Record* 51: 522–538.

Strathern, A.J. 1971. *The rope of moka*. Cambridge: Cambridge University Press.

Young, M.W. 1971. *Fighting with food: Leadership, values and social control in a Massim society*. Cambridge: Cambridge University Press.

Gilbert, Milton (1909–1979)

Irving B. Kravis

Gilbert was born in Philadelphia and educated in that city, receiving his doctorate from the University of Pennsylvania. His contributions to economics were both in the statistical and substantive domains.

His work as an economic statistician occupied the early and middle years of his career and was centred on the development of national accounts. His first major position was in the US Department of Commerce as Editor of the *Survey of Current Business.* He became chief of the national income division in the Commerce Department in 1941 and for the next 10 years presided over the development of the US system of national income and product accounts. In 1951 he went to the Organization for European Economic Cooperation (OEEC) first as head of the Statistics and National Accounts Division and then as head of the combined Economics and Statistics Division. At the OEEC he initiated and co-authored the first systematic international comparison of national products and currency purchasing powers. His growing attention to substantive economic issues was reflected in his leading role in an influential OEEC report on the wage–price spiral.

In 1960 Gilbert went to the Bank for International Settlements (BIS) as Economic Adviser and later headed the Bank's Monetary and Economic Department. In addition to his analyses of world monetary problems, much of it focused on the workings and the breakdown of the Bretton Woods system, Gilbert greatly strengthened the statistical work of the BIS especially with regard to the collection and analysis of data on the Eurocurrency market.

See Also

► International Income Comparisons
► Social Accounting

Selected Works

1947. (With others.) *U.S. national income supplements*. Washington, DC: US Government Printing Office.
1954. (With I.B. Kravis.) *An international comparison of national products and the purchasing power of currencies: A study of the United States, the United Kingdom, France, Germany and Italy*. Paris: Organization for European Economic Co-operation.
1968. *The gold–dollar system: Conditions of equilibrium and the price of gold*. Princeton: International Finance Section, Princeton University.
1980. *Quest for world monetary order: The gold–dollar system and its aftermath*. Posthumous, ed. P. Oppenheimer and M.G. Dealtry. New York: Wiley.

Gilman, Charlotte Perkins (1860–1935)

B. Berch

Keywords
Gilman, C. P.; Women's work and wages

JEL Classifications
B31

Gilman was born on 3 July 1860 in Hartford, Connecticut, and died on 17 August 1935 in Pasadena, California. Known worldwide as a feminist theorist and a generally iconoclastic social critic, Gilman was a major intellectual force in America at the turn of the 20th century. Largely self-educated, problems with her first marriage led her to separate from her first husband and begin an unconventional freelance

life based in California, earning her living from her lecturing and writing. *Women and Economics* (1898) was her first book-length exposition of her theory of the evolution of gender relations. Influenced by the ideas of Edward Bellamy, Lester Frank Ward, Darwin, the Webbs and G. Bernard Shaw, she explained that human institutions (like the species itself) has evolved over time, favouring the survival of the best adapted. A major exception, however, was the definition of 'women's place'. Here social development had been frozen by Tradition. Women were confined to households which were no longer the locus of any socially productive activity, since now the factory produced the needed consumption goods, and children were better raised in schools, by professionals. The role of full-time housewife and mother had become anachronistic, reducing women to the state of social parasites. As she also argued in her 1903 classic *The Home* that, for their own progress and for the progress of human civilization overall, women would have to leave these domestic prisons and take up socially useful work in the larger world of production. In the articles and didactic fiction that she wrote for her monthly magazine the *Forerunner*, she developed a wide range of startlingly rational ideas for social reorganization.

Selected Works

1892. The yellow wall-paper. *New England Magazine.* January. Reprinted in *The Charlotte Perkins Gilman reader*, ed. A. Lane. New York: Pantheon, 1980.
1898. *Women and economics*, ed. C. Degler. New York: Harper & Row, 1966.
1903. *The home.* Introduction by W. O'Neill. Urbana: University of Illinois Press, 1972.
1915. Herland. *The Forerunner* 6. Reprinted, with introduction by A. Lane, as *Herland: A lost feminist Utopian novel*. New York: Pantheon, 1978.

Gini Ratio

Camilo Dagum

JEL Classifications
D3

In a national economy, the price system determines both resource allocation and the income distribution. The imputation to the factors of production of the mass of income associated with an economy's output determines its distribution by factor shares, or *functional* income distribution. This mainstream of research follows Ricardo's (1817) contribution. Another mainstream of research was initiated by Pareto (1895, 1897), and deals with the distribution of a mass of income among the members of a set of economic units (family, household, individual), considering either the total income of each economic unit or its disaggregation by source of income, such as wages and salaries, property income, self-employment income, transfers, etc. This type of inquiry deals with distribution by size of income, or *personal* income distribution, and the quantitative assessment of the relative degree of income inequality among the members of a given set of economic units. Such inquiries provide basic quantitative information in support of a comprehensive research strategy on income distributions, including causal explanations for social welfare and policy.

It is of interest to remark that Pareto's research on income distribution was motivated by the polemic he engaged in with French and Italian socialists concerning the ways and means of achieving a less unequal distribution. Thus, the actual *measurement* of inequality was brought to the fore, with its main purposes the assessment of (i) the evolution of inequality in a given country or region, and (ii) the relative degree of inequality between countries or regions.

In a series of methodological and applied contributions Corrado Gini (1955) enriched this field

of research. In 1910 he corrected the interpretation of Pareto's inequality parameter and, in 1912, proposed a new measure of income inequality, the Gini ratio.

Pareto (1896, 1897) specified three versions of this model of income distribution. The most widely used model is Pareto Type I

$$S(x) = 1 - F(x) = (x/x_0)^{-\alpha}, \quad 0 < x_0 < x, \alpha > 1, \tag{1}$$

where $S(x) = P(X > x)$ is the survival distribution function (SDF) of the income variable X, $F(x)$ is the cumulative distribution function (CDF), x_0 is the minimum value of X, α is a scale-free inequality parameter, and the mathematical expectation of income is

$$\mu = E(X) = \alpha x_0^{\alpha} \int_{x_0}^{\infty} x^{-\alpha} dx = \alpha x_0/(\alpha - 1). \tag{2}$$

Pareto seems to have assumed that income growth implies less income inequality. This assumption, together with eqn (2), led him to the conclusion that income inequality is an increasing function of α. Gini (1910) reversed this interpretation, proving that, given model (1), income inequality is a decreasing function of α. Gini's rationale was as follows: given n units with incomes $x_1 \le x_2 \le x_3$, . . . , $\le x_n$, the average of the last $m (m \le n)$ income units $\Sigma_{i=0}^{m-1} x_{n-i}/m$ is greater than or equal to the average income $\mu = \Sigma_{i=0}^{n} x_i/n$ of the population, hence, there exists a $\delta \ge 1$ such that

$$\left(\sum_{i=0}^{m-1} x_{n-i} / \sum_{i=1}^{n} x_i \right)^{\delta} = m/n, \qquad \delta \ge 1, \tag{3}$$

Equation (3) is known as the Gini model. Gini (1910) interpreted the scale-free parameter δ as a measure of income inequality and called it a *concentration ratio* because it is an increasing function of the concentration of income in the upper income groups. For this reason, Gini called eqn (3) a *concentration curve,* where the abscissa represents the CDF $F(x_m) = m/n$ and the ordinate the income share $\sum_{i=1}^{m} x_i / \sum_{i=1}^{n} x_i$, $m = 1$, 2 , . . . , n , δ being an unknown parameter that has to be estimated.

Using the CDF $F(x)$ and the Lorenz curve $L(x)$ (also called the Lorenz-Gini curve since it was independently introduced by both authors), eqn (3) takes the form

$$1 - F(x) = [1 - L(x)]^{\delta}, \qquad \delta \ge 1, \tag{4}$$

where

$$L(y) = (1/\mu) \int_{x}^{y} x \mathrm{d}F(x). \tag{5}$$

Replacing $F(x)$ from model (1) into eqns (4) and (5), Gini (1910) proved that $\delta = \alpha/(\alpha - 1)$ and thus reversed Pareto's interpretation of α. In fact, when $\alpha \to \infty$, $\delta \to 1$ and $F(x) = L(x)$, and the mass of income is equally distributed.

Gini (1912) specified the Gini mean difference with and without replacement. The latter is by definition

$$\Delta = \sum_{j=1}^{n} \sum_{i=1}^{n} |x_j - x_i|/n(n-1),$$

$$0 \le \Delta \le 2\mu, \tag{6}$$

and using the Riemann-Stieltjes integral, which covers, as particular cases, both discrete and continuous distributions, we have

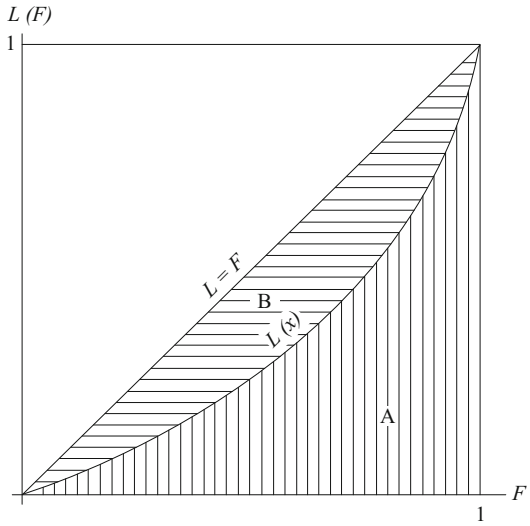
$$\Delta = \int_0^{\infty} \int_0^{\infty} |y - x| \mathrm{d}F(x) \mathrm{d}F(y), \tag{7}$$

where X and Y are identically and independently distributed variables. When $x_1 = x_2 =$ . . . $= x_n$, $\Delta = 0$, and when $x_1 = x_2 =$ . . . $= x_{x-1} = 0$ and $x_n = n\mu$ (the total income), $\Delta = 2\mu$.

Since Δ is a monotonie increasing function of the degree of income inequality, Gini (1912) specified

$$G = \Delta/2\mu, \qquad 0 \le G \le 1 \tag{8}$$

G



Gini Ratio, Fig. 1 Lorenz curve *L(x)* and Gini ratio *G*

as an income inequality measure. Equation (8) is known as the Gini ratio or Gini index and it is widely used in theoretical and applied research on income and wealth distributions.

Gini (1914) proved the important theorem that $G = \Delta/2\mu$ is equal to twice the area between the equidistribution line $F(x) = L(x)$ and the Lorenz curve $L(x)$ (see Fig. 1). Moreover,

$$\begin{aligned} G = \Delta/2\mu &= 2\int_0^1 (F - L)\mathrm{d}F \\ &= (2/\mu)\int_0^\infty x\left[F(x) - \frac{1}{2}\right]\mathrm{d}F(x) \\ &= (2/\mu)\int_0^\infty x\left[\frac{1}{2} - S(x)\right]\mathrm{d}F(x). \end{aligned} \tag{9}$$

For the discrete case, it follows from eqns (6) and (8), that

$$\begin{aligned} (F(x)) &= (1/\mu)\int_0^x y\mathrm{d}F(y)G = 2B = 1 - 2A = 1 - 1\int_0^1 L\mathrm{d}F \\ G &= [2/n(n-1)\mu]\sum_{k=1}^{n} kx_k - (n+1)/(n-1) \\ &= (n+1)(n-1) - [2/n(n-1)\mu]\sum_{k=1}^{n}(n-k+1)x_k, \end{aligned} \tag{10}$$

showing that the welfare function underlying the Gini ratio is a rank-order-weighted sum of the economic units' income shares.

The properties that an income inequality measure must fulfil were first discussed by Dalton (1920). It can be shown (Dagum 1983, pp. 34–5) that G fulfils the properties of (i) transfer, (ii) proportional addition to incomes, (iii) equal addition to incomes, (iv) proportional addition to persons, (v) symmetry, (vi) normalization, and (vii) operationality.

The Gini ratio is sensitive to transfers to all income levels. In fact, it follows from eqn (10), that a transfer of h dollars from the richer j to the poorer i theorem, without modifying their income ranks, is

$$\Delta G(j, i; h) = -2(j - i)h/n(n-1)\mu > 0, \qquad j > i, \tag{11}$$

therefore $-\Delta G$ is an increasing function of $j - i = F(x_j) - F(x_i)$ and a decreasing function of both n and μ. The maximum reduction of G is achieved when $h = (x_j - x_i)/2$, and is not necessarily given by eqn (11) unless the transfer fulfils certain conditions with respect to the original income ranking of the population.

Often, the Gini ratio is misinterpreted when it is incorrectly claimed that it attaches more weight to transfers to income near the mode of the distribution than at the tails. In particular, the misinterpretation arises when eqn (11) instead of eqn (9) is applied to unimodal distributions when assessing the relative sensitivity of G to income transfers. Consequently, the assumptions supporting the mathematical structure of eqn (11) are ignored.

It follows from eqn (10) that the Gini ratio fulfils the duality principle between the representation of an inequality measure (I) satisfying the principle of transfer, i.e. $I = E[V(x)]$, and that of a social welfare (SW) function, i.e. $SW = E[-V(x)]$, where $-V(x)$ is concave, or more generally, S-concave (Berge 1966). It follows from eqn (9), that two equivalent forms of $V(x)$ in $G = E[V(x)]$ are

$$V(x) = 2xF(x)/\mu - 1, \quad \text{and} \quad V(x) = x[2F(x) - 1]/\mu. \tag{12}$$

Sen (1974) introduced an axiomatic system for the SW interpretation of the Gini ratio based on the individual income ranking of the population suggested by the structure of eqn (10). Following Sen's ideas, Kakwani (1980, pp. 77–9) presented a SW interpretation of the Gini ratio as a function of income. Both approaches can be presented in a compact form by making use of the SDF $S(x) = 1 - F(x)$ and the first moment survival distribution function $S_1(x) = 1 - L(x)$. In fact, specifying the SW function

$$\mathrm{SW}(X) = E[Xv(x)] \tag{13}$$

where $v(X)$ is a decreasing and differentiable function of X, and making $v(X) = 2S(X) = 2(1 - F(X)]$, i.e., twice the frequency of economic units with income greater than X, we deduce

$$\mathrm{SW}(X) = 2\int_0^\infty xS(x)\ \mathrm{d}F(x) = \mu(1 - G), \tag{14}$$

which proves Sen's (1974, p. 410) theorem that the SW function (14) ranks a set of distributions of a constant total income and population in precisely the same way as the negative of the Gini ratio of the respective distributions, i.e. in reverse order from that by the cardinal value of the Gini ratio. On the other hand, making $v(X) = bS_1\ (X) = b[1 - L(X)]\ ,\ b > 0$ and $\int_0^\infty v\ (x)\ \mathrm{d}F(x) = 1$, where $S_1(x)$ is the income share of the economic units with income greater than x, we deduce

$$b\int_0^\infty [1 - L(x)]\mathrm{d}F(x) = b(1 + G)/2 = 1, \text{and}$$
$$\mathrm{SW}(X) = [2/(1 + G)]\int_0^\infty x[1 - L(x)]\mathrm{d}F(x) = \mu/(1 + G), \tag{15}$$

which also states that the SW function (15) is a decreasing function of the Gini ratio. The result obtained in eqn (14) supports Sen's (1976, p. 384) cogent statement that 'one might wonder about the significance of the debate on the non-existence of any additive utility function which ranks income distributions in the same order as the Gini ratio'.

The Gini ratio stimulated important contributions such as:

(i) The construction of a confidence interval for G. Given a random sample of size n, eqn (10) is an unbiased estimator of G. However, income distribution data are presented by class intervals, hence Gini (1914) proposed the formula $\mathrm{G}_L = 1–2A$, where A is the area under the Lorenz curve (Fig. 1) estimated by application of the trapezoidal approximation to

$$\int_0^1 L\ dF,$$

thus underestimating G because the trapezoidal rule implies that within each interval, income is equally distributed. Gastwirth (1972) derived an upper bound G_u, by maximizing the spread within each income interval, and proposed (G_L, G_u) as a confidence interval within which a parametric estimate of G should fall. Dagum (1980a) proved that his confidence interval is a necessary but not sufficient condition to assess a model goodness of fit.

(ii) The Gini ratio gives a welfare ranking (weak ordering) of a set of income distributions of a constant mass of income and over a constant population, and a strict partial ordering among the subset of income distributions with non-intersecting Lorenz curves. This conclusion is further supported by eqns (14) and (15).

(iii) The welfare ranking of income distributions with equal and different means can be obtained via a decision function $R(G, D)$, where the ratio G states the preference for less inequality (inequality aversion) regardless of the mean income (so that the partial derivative $R_G < 0$), and the relative

economic affluence D (Dagum, 1980b, 1987) states the preference for more income (poverty aversion), so that the partial derivative $R_D > 0$.

(iv) Research on the economics of poverty led Sen (1976) to the specification of an axiomatic structure of a new poverty measure as a function of (a) the relative frequency of the poor members of the population, (b) a weighted average of the poverty gap, i.e. the aggregate shortfall from the poverty line of the poor population, and (c) the Gini ratio of the income distribution of the subpopulation with incomes below the poverty line.

(v) Gini (1932) introduced a new coordinate system taking as the abscissa the egalitarian line $F = L$ and as the ordinate the distance between the Lorenz curve and the egalitarian line. Gini thoroughly analysed this new coordinate system and its relation to the G ratio. Kakwani (1980, ch. 7) worked with a similar transformation.

(vi) Analysing consumer behaviour in India, Mahalanobis (1960) extended and generalized the Lorenz curve and the Gini ratio with the introduction of the concentration curve and ratio, respectively. Other authors such as Kakwani (1980, chs 8–14) made further contributions and dealt with the relationships among the distribution of several economic variables such as expenditures and income after tax, and investigated the degree of tax and public expenditure progressivity or regressivity. If $y = g(x)$ is the function of income that is the object of inquiry, $g(x)$ must be non-negative. For the particular case of $g(x) = x$, the concentration curve and ratio are identical to the Lorenz curve and Gini ratio, respectively. Moreover, if $g(x)$ is an increasing and differentiable function of x, i.e. $g'(x) > 0$, then the concentration ratio is equal to the Gini ratio for the function $g(x)$.

(vii) The decomposition approach disaggregates a population according to some relevant socio-economic attributes and analyses the equality within each subpopulation and between them, and assesses the contribution of each subpopulation to overall inequality. This approach also disaggregates the income variable by source of income such as wages and salaries, self-employment, pension and government transfers. Bhattacharya and Mahalanobis (1967) were the first to deal with the decomposition of the Gini ratio. Several authors made further contributions to this topic, among them Pyatt (1976) and Shorrocks (1983).

Bibliography

Berge, C. 1966. *Espaces topologiques, fonctions multivoques*, 2nd edn. Paris: Dunod.

Bhattacharya, N., and P.C. Mahalanobis. 1967. Regional disparities in household consumption in India. *Journal of the American Statistical Association* 62: 143–161.

Dagum, C. 1980a. The generation and distribution of income, the Lorenz curve and the Gini ratio. *Economie Appliquée* 33 (2): 327–367.

Dagum, C. 1980b. Inequality measures between income distributions with applications. *Econometrica* 48 (7): 1791–1803.

Dagum, C. 1983. Income inequality measures. In *Encyclopedia of statistical sciences*, ed. S. Kotz and N.L. Johnson, vol. IV, 34–40. New York: Wiley.

Dagum, C. 1987. Measuring the economic affluence between populations of income receivers. *Journal of Business and Economic Statistics* 5 (1): 5–12.

Dalton, H. 1920. The measurement of the inequality of incomes. *Economic Journal* 30 (September): 348–361.

Gastwirth, J.L. 1972. The estimation of the Lorenz curve and Gini index. *Review of Economics and Statistics* 54: 306–316.

Gini, C. 1910. Indici di concentrazione e di dependenza. *Atti delta III Riunione delta Società Italiana per il Progresso delle Sciencze*, in Gini (1955), 3–120.

Gini, C. 1912. Variabilitàe mutabilità. *Studi economico-giuridici, Università di Cagliari* III, 2a, in Gini (1955), 211–382.

Gini, C. 1914. Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti,* in Gini (1955), 411–459.

Gini, C. 1932. Intorno alle curve di concentrazione. *Metron* 9(3–4), in Gini (1955), 651–724.

Gini, C. 1955. *Memorie di metodologia statistica.* Vol. 1: *Variabilità > concentrazione,* ed. E. Pizetti and T. Salvemini. Rome: Libreria Eredi Virgilio Veschi.

Kakwani, N. 1980. *Inequality and poverty: Methods of estimation and policy applications*. Oxford: Oxford University Press.

Mahalanobis, P.C. 1960. A method of fractile graphical analysis. *Econometrica* 28 (2): 325–351.
Pareto, V. 1895. La legge della domanda. *Giornale degli Economisti* 5: 59–68.
Pareto, V. 1896. Ecrits sur la courbe de la répartition de la richesse. In *Oeuvres complètes de Vilfredo Pareto*, ed. Giovanni Busino. Geneva: Librairie Droz, 1965.
Pareto, V. 1897. *Cours d' éeconomie politique,* ed. G.H. Bousquet and G. Busino. Geneva: Librairie Droz, 1964.
Pyatt, G. 1976. On the interpretation and disaggregation of Gini coefficients. *Economic Journal* 86 (June): 243–255.
Ricardo, D. 1817. Principles of political economy. In *Works and correspondence of David Ricardo*, ed. P. Sraffa, vol. I. Cambridge: Cambridge University Press. 1951.
Sen, A.K. 1974. Information bases of alternative welfare approaches. *Journal of Public Economics* 3: 387–403.
Sen, A.K. 1976. Poverty: An ordinal approach to measurement. *Econometrica* 44 (2): 219–231.
Shorrocks, A.F. 1983. The impact of income components on the distribution of family income. *Quarterly Journal of Economics* 98 (2): 311–326.

Gini, Corrado (1884–1965)

Camilo Dagum

Keywords

Demography; Distribution of income and wealth; Gini coefficient; Gini identity; Gini mean difference; Gini, C.; Human capital; Index numbers; Inequality (measurement); Internal migration; International migration; Statistics and economics

JEL Classifications

B31

Gini, perhaps best known to economists because of the Gini Coefficient, was born in Motta di Livenza, Italy and died in Rome. He studied at the University of Bologna; his doctoral thesis *Il sesso dal punto di vista statistico* (1908), defended in 1905, was awarded the Vittorio Emanuele prize for social sciences. Gini distinguished himself as a teacher and a researcher. In 1909 he was appointed an assistant professor of the University of Cagliari, becoming full professor a year later. Gini won a chair at the University of Padova in 1913, then joined the University of Rome in 1925, where in 1955 he was awarded the distinction of emeritus professor. Social scientist and statistician, Gini taught economics, statistics, sociology and demography, making path-breaking contributions to these highly related disciplines. Among them we mention the neo-organicist theory (Gini 1909, 1924a) that presents a dynamic theory of society in which demographic factors (differential birth rates among social classes and social mobility) play a basic role. In this theory, Gini introduced and analysed self-conservation, self-regulative and self-re-equilibrating mechanisms, thus offering a well-structured anticipation of Wiener's cybernetics, von Bertalanffy's general system theory and modern disequilibrium economics. He provided new insights to the analysis of inter- and intra-national migrations (Gini 1948) and demographic dynamics (Gini 1908, 1909, 1912a, 1931). He developed a methodology to evaluate the income and wealth of nations (Gini 1914a, 1959) including a discussion of human capital, already present in his research on the causes and consequences of international migrations. In this context he specified a model of income and wealth distributions and a measure of income and wealth inequalities (Gini 1909, 1912b, 1914b, 1955). Gini's research interests motivated important contributions to statistics and economics, such as the Gini identity (1921, 1924b) on price index numbers, the Gini mean difference (1912b), the transvariation theory (Gini 1916, 1960), the index of dissimilarity (Gini 1914c) and the Gini Coefficient. Gini founded several scientific journals, such as *Metron* and *Genus*, and academic institutions, such as the Institute and Faculty of Statistics, Demography and Actuarial Sciences of the University of Rome; and was the organizer and first president (1926–1932) of the Istituto Centrale di Statistica. An extraordinarily prolific writer and thinker, endowed with powerful new ideas that he developed in more

than 70 books and 700 articles, Gini was in the 20th century a true Renaissance man.

Selected Works

1908. *Il sesso dal punto di vista statistico*. Milan: Sandrom.
1909. Il diverso accrescimento delle classi sociali e la concentrazione della ricchezza. *Giornale degli Economisti* 38: 27–83.
1912a. *I fattori demografici dell'evoluzione delle nazioni*. Turin: Bocca.
1912b. Variabilità e mutabilità. Reprinted in Gini (1955).
1914a. *L'ammontare e la composizione della ricchezza delle nazioni*. 2nd ed., Turin: UTET, 1962.
1914b. Sulla misura della concentrazione e della variabilità dei caratteri. Reprinted in Gini (1955).
1914c. Di una misura della dissomiglianza tra due gruppi di quantità e delle sue applicazioni allo studio delle relazioni statistiche. *Atti del R. Istituto Veneto di Scienze, Lettere ed Arti* 74, Part II.
1916. Il concetto di transvarizione e le sue prime applicazioni. Reprinted in Gini (1960).
1921. Sull'interpolazione di una retta quando i valori della variabile indipendente sono affetti da errori accidentali. *Metron* 1:63–82.
1924a. *Patologia economica*, 5th ed. Turin: UTET, 1954.
1924b. Quelques considérations au sujet de la construction des nombres indices des prix et des questions analogues. *Metron* 4:3–162.
1931. *Le basi scientifiche della politica della popolazione*. Catania: Studio Editoriale Moderno.
1948. Apparent and real causes of American prosperity. *Banca Nazionale del Lavoro Quarterly Review* 6, July–September.
1955. *Memorie di metodologia statistica*. Vol. 1: *Variabilità e concentrazione*, ed. A. Pizzetti and T. Salvemini. Rome: Veschi.
1959. *Ricchezza e reddito*. Turin: UTET.
1960. *Transvariazione*, ed. G. Ottaviani. Rome: Libreria Goliardica.

Gioia, Melchiorre (1767–1829)

U. Pagano

Born in Piacenza, Italy, Melchiorre Gioia actively participated in the turbulent political life of his time, ending up in prison more than once.

Gioia became a Catholic priest; however, his gospel was that man should obtain the 'maximum product with the minimum expenditure of effort'. This 'principle' inspires his main contribution to the development of economic analysis: the principle of the association and of division of work. The principle of the association of work states that the cooperation of qualitatively equal kinds of labour increases efficiency in production. Cooperation has considerable advantages even if it is not associated with the division of labour (i.e. the association of qualitatively different kinds of labour). As to the nature and the advantages of the division, Gioia differentiates himself from Smith on two points. Firstly, he maintains that the division of labour (as well as undifferentiated cooperation) can and does exist in the animal and human world independently of exchange. Secondly, Gioia relates the advantages of the division of labour more to the saving of natural or acquired human skills than to the creation of job specific skills. Natural skills are saved if the individuals possessing them dedicate themselves only to those occupations which require them while other people perform the remaining activities. As to the scope of acquired skills, specialization saves training time. The more pronounced the specialization of the members of an association, the narrower the set of tasks which they have to learn.

Melchiorre Gioia criticized the laissez-faire policies advocated by the English school of political economy. He anticipated some aspects of the theories of market failures due to externalities and monopolies and favoured state intervention for correcting them. He also advocated protectionistic policies aimed at the development of infant industries.

Gioia's greatest work is *Il Nuovo Prospetto delle Scienze Sociali* (1815), the first volume of which contains his principle of the association and division of labour. His views on state intervention are well expressed in *Discorso popolare sulle manifatture nazionali e tariffe daziarie dei commestibili ed il caro prezzo del vitto* (1802). His other works include: *Del merito e delle ricompense*, etc. (1818), *Filosofia della Statistica* (1826), *Indole, estensione e vantaggi della Statistica*, (1809), *Logica Statistica* (1808) and *Tavole Statistiche* etc. (1808).

An exhaustive examination of the economic theory of Melchiorre Gioia can be found in Barucci (1965).

Selected Works

1815. Il nuovo prospetto delle scienze sociali. Lugano.

References

Barucci, P. 1965. *Il pensiero economico di Melchiorre Gioia*. Milan: Giuffrè.

GIS Data in Economics

Henry G. Overman

Abstract

Geographical information systems (GIS) are used for inputting, storing, managing, analysing and mapping spatial data. This article considers the role each of these functions can play in economics. GIS can map economic data with a spatial component, generate additional spatial data as inputs to statistical analysis, calculate distances between features of interest and define neighbourhoods around objects. GIS also introduce economics to new data. For example, remote sensing provides large amounts of data on the earth's surface. These data are of inherent interest, but can also provide an exogenous source of variation and allow the construction of innovative instrumental variables.

Keywords

Geographical information systems (GIS) and economics; Spatial data; Hedonic analysis

JEL Classifications

R12

Geographical information systems (GIS) are used for inputting, storing, managing, analysing and mapping spatial data. In this article, we consider each of these functions to help assess the role that GIS can play in economic analysis. Of course, a wide range of software can provide similar functions for quantitative data, so it is the geographical, or spatial, element that separates GIS. That spatial dimension is the focus here. One important aspect of GIS that is not covered is the choice of software. Standard texts, such as Longley et al. (2005, ch. 7), consider the question of appropriate software in some depth.

At the outset, note that, while GIS are widely used in business, government and a range of academic disciplines, their application in economics has to date been more limited. The most frequent application in economics is the use of GIS to visualize or map economic data with a spatial component. Most entry-level courses in econometrics begin with a plea to 'plot the data' at an early stage of the analysis to help identify trends, outliers and so forth. Much the same could be said of the role of mapping spatial data, and GIS provide a simple and efficient way to do this.

Less common, but arguably more interesting, is the use of GIS storage and management functions to generate additional data as inputs into further statistical analysis. In the simplest case this will involve using GIS to manage spatial data from a variety of sources. Many of these sources – for example, sampling and census data – will be familiar to economists, others, such as aerial photography and remote sensing

data, less so. The spatial nature, or format, of the data will depend on the geographical data model used. The two most common models are raster format (assigning a code to each cell on a regular grid) and vector format (assigning a code to, and providing coordinates for, irregular polygons). GIS provide tools for moving between these different geographical data models. While the methods used to do this are rather intuitive, the devil is in the detail. As with the implementation of pre-packed econometric routines, one should understand the underlying basis of these transformations before proceeding. These issues are covered at depth in most, if not all, of the standard references, and we do not consider them further here.

More generally, it is the ability of GIS to reconcile spatial data from different sources that allows the creation of new data-sets. In the simplest case, this may involve combining socio-economic data from different spatial units – for example, population data from US census tracts with employment data for US zip codes. Many economists will be used to using ready-made concordances (that is, mappings from one set of spatial units to another) for undertaking such data merges. GIS bring the flexibility of allowing users to define their own concordances between different geographical units of observation when faced with data from different sources.

The construction of more ambitious data-sets is possible if one is willing to draw on a range of analytical functions available in the more advanced GIS. GIS can be used to identify whether observations occur at particular locations and, if so, to identify the characteristics of observations at those locations. For example, one of the most frequent applications of GIS in economics to date has been to identify and characterize properties for use in hedonic analysis (see Bateman et al. 2002, for a review). At its most basic, this will simply involve the merging of different data-sets as described above. However, much more complex analysis is possible. Given that GIS data are spatial, a natural use is to measure the distance between observations or between observations and other features of interest. These distances could be physical distances or network distances (for example, along a transport network), or involve some more general concept of social distance.

Observation-to-observation distance calculations have been widely applied in the fields of biology and biomedical sciences through the statistical analysis of spatial point patterns (see Diggle 2003). Knowing the distance between observations is useful if we think that there may be interactions between them and that the strength of these interactions is mitigated by distance. For example, in industrial organization models of spatial competition the intensity of competition may depend on the distance between firms. Observation-to-observation distances are also useful when the underlying entities are free to choose their location and we want to assess whether there are systematic patterns to those location decisions. For example, the study of localization asks whether firms in a particular industry tend to be spatially concentrated relative to overall economic activity. If they do, one would expect the observation-to-observation distances to be less for firms in that industry than for a randomly chosen set of firms from the economy at large. The increasing availability of geo-referenced economic data suggests that the application of appropriately adapted procedures will become more common in economics (see Duranton and Overman 2005). Hedonic analysis again provides the most frequent application of GIS to calculate distances of observations from other features. For example, in their study valuing rail access Gibbons and Machin (2005) use GIS to measure the proximity of properties to rivers, coasts, woodlands, roads, railway lines and airports.

In addition to the calculation of distances, GIS can be used to construct measures of area or to define neighbourhoods (or 'buffers') around objects. For example, Burchfield et al. (2006) in their study of urban sprawl use GIS to calculate the percentage of the urban fringe – defined as a 20-kilometer buffer around existing development – that lies above water-yielding aquifers.

These examples cover the main types of spatial analyses that are undertaken to construct spatial data in economic applications, but others are

possible and should be covered in any of the standard texts. It should be noted that in advanced GIS these operations can be done both interactively and automatically using batch files (that is, where the user writes a sequence of commands in a file that the computer implements one by one). Both approaches, but particularly the latter, involve fairly large fixed costs in terms of both purchasing software and learning how to implement the relevant procedures. There are other methods for conducting many of these analyses that do not imply the use of GIS. For example, great circle distances can easily be calculated on the basis of latitude and longitude (see Overman and Ioannides 2004). Whether the fixed cost investment is worthwhile will depend on individual circumstances. The benefits can be substantial. In many circumstances, GIS calculations should be more accurate than short cuts implemented with the use of non-spatial software, and some analysis such as the calculation of areas and buffers is much easier to implement in GIS.

GIS also introduce economics to new sources of data. In particular, remote sensing from either satellite or aerial photography, or digitized geological maps, can provide a huge amount of data on the earth's surface. Early applications using these kinds of data tended to focus on issues arising from natural resource management such as valuing timber yields from forested areas. However, data on land cover and land use (that is, the physical features that cover the land and what those features are used for), soil type, geological and landscape features, elevation and climate are opening up new avenues of research. These data sources allow the description of different features of the economic landscape that one might seek to explain. For example, Burchfield et al. (2006) use remote sensing data to track the evolution of land use on a grid of 8.7 billion 30 × 30 metre cells covering the conterminous United States and then seek to explain differences in land development patterns across cities. Another example is Rappaport's (2006) study of the role that weather plays in explaining population changes in US counties. The meteorological data that he uses comes from 6,000 meteorological stations and covers 20 winter, summer and precipitation variables. GIS analysis by the Spatial Climate Analysis Surface at Oregon State University applied to this meteorological data allows the construction of weather variables for a two-kilometre grid covering the continental United States.

GIS data also have the potential to contribute to a range of established areas of study, particularly because data on the earth's surface can provide an exogenous source of variation and thus allow researchers to construct instrumental variables using GIS. Some examples should help to make this idea concrete. Hoxby (2000) is interested in whether competition among public schools improves school outcomes. That is, do cities with more school districts have better public schools and less private schooling? The problem that the analysis needs to confront is that, for a city of a given size, better public schools and fewer private schools in a city should imply more school districts. That is, the number of districts is endogenous to public school quality. What is needed is an instrument that should determine the supply of school districts but that is independent of the local public school quality. Hoxby argues that the number of streams in a metropolitan area provides such an instrument. Cities with a large number of streams end up with more school districts for reasons that are surely nothing to do with public school quality. Hoxby provides a well-known example of the strategy, although not of the use of GIS, as her work is based on the study of detailed paper maps.

Rosenthal and Strange (2005) provide an example of the use of GIS to implement such an instrumental variables strategy. They are interested in whether density of employment helps determine wages. The problem is that higher wages should attract more workers and lead to higher employment densities. That is, density may be caused by wages and not vice versa. Rosenthal and Strange argue that the density of employment will be partly determined by the height of buildings in a location. They point out that the height of buildings is, in turn, partly dependent on the underlying geology of the site. Given that geology should not determine wages directly (they are studying cities, not agricultural

production), the underlying geology can be used as an instrument. Locations with a suitable underlying geology can have higher buildings and higher employment density, and should thus have higher wages. Rosenthal and Strange use GIS data on the type of underlying bedrock, seismic and landslip hazard as instruments for the density of employment in their regressions of wages on employment density. Such examples suggest a potentially important role in future work for GIS data as a component in novel instrumentation strategies.

This piece has only skimmed the surface of the potential applications of GIS in economics. As spatially referenced socio-economic data becomes more widely available, it is to be expected that the scope for applications can only increase.

See Also

- ▶ Location Theory
- ▶ New Economic Geography
- ▶ Spatial Econometrics
- ▶ Spatial Economics

Bibliography

Bateman, I., and A. Lovett. 1998. Using geographical information systems (GIS) and large area databases to predict yield class: A study of Sitka spruce in Wales. *Forestry* 71: 147–168.

Bateman, I., A. Jones, A. Lovett, I. Lake, and B. Day. 2002. Applying geographical information systems (GIS) to environmental and resource economics. *Environmental and Resource Economics* 22: 219–269.

Burchfield, M., H. Overman, D. Puga, and M. Turner. 2006. The determinants of sprawl: A portrait from space. *Quarterly Journal of Economics* 135: 587–633.

Diggle, P. 2003. *Statistical analysis of spatial point patterns*. London: Arnold.

Duranton, G., and H. Overman. 2005. Testing for localisation using micro-geographic data. *Review of Economic Studies* 72: 1077–1106.

Gibbons, S., and S. Machin. 2005. Valuing rail access using transport innovations. *Journal of Urban Economics* 57: 148–169.

Hoxby, C. 2000. Does competition among public schools benefit students and taxpayers? *American Economic Review* 90: 1209–1238.

Longley, P., M. Goodchild, D. Maguire, and D. Rhind. 2005. *Geographical information systems and science*. 2nd ed. Chichester: John Wiley and Sons.

Overman, H., and Y. Ioannides. 2004. The spatial evolution of the US urban system. *Journal of Economic Geography* 4: 131–156.

Rappaport, J. 2006. Moving to nice weather. Working paper no. 03–07. Research Division, Federal Reserve Bank of Kansas City.

Rosenthal, S. and Strange, W. 2005. The attenuation of agglomeration economies: A Manhattan skyline approach. Working paper. University of Toronto.

Global Analysis in Economic Theory

Steve Smale

Abstract

Global analysis in economics puts the main results of classical equilibrium theory into a global calculus context. The advantages of this approach are: (*a*) the proofs of existence of equilibrium are simpler (the main tool is the calculus of several variables); (*b*) comparative statics is integrated into the model in a natural way; (*c*) the calculus approach is closer to the older traditions of the subject; and (*d*) as far as possible the proofs of equilibrium are constructive.

Keywords

Arrow–Debreu theory; Calculus of variations; Comparative statics; Existence of equilibrium; Global analysis in economic theory; Inverse function th; Proofs of equilibrium; Sard's th; Walras's Law

JEL Classifications

D5

The goal here is to illustrate 'global analysis in economics' by putting the main results of classical equilibrium theory into a global calculus context. The advantages of this approach are fourfold:

1. The proofs of existence of equilibrium are simpler. Kakutani's fixed point theorem is not used, the main tool being the calculus of several variables.
2. Comparative statics is integrated into the model in a natural way, the first derivatives playing a fundamental role.
3. The calculus approach is closer to the older traditions of the subject.
4. In so far as possible the proofs of equilibrium are constructive. These proofs may be implemented by a speedy algorithm, which is Newton's method modified to give global convergence. On the other hand, the existence proofs are sufficiently powerful to yield the generality of the Arrow–Debreu theory.

Only two references are given at the end of this entry, each containing an extensive bibliography with historical notes. The two references themselves give detailed, expanded accounts of the subject of global analysis in economic theory.

Let us proceed to an account of this model. The basic equation of equilibrium theory is 'supply equals demand', or in symbols, $S(p) = D(p)$. Since we are in a situation of several markets, there are several variables in this equation. Equilibrium prices are obtained by setting the excess demand $z = D - S$ equal to zero and solving. Consider this function z on a more abstract level.

Suppose that, given an economy of l markets, or of l commodities with corresponding prices written as $p_1,\ldots, p_l$, the excess demand for the ith commodity is a real valued function $z_i = z_i, (p_1,\ldots, p_l)$ $p_j \geq 0$ and we form the vector $z = (z_l,\ldots, z_1)$. Thus the excess demand can be interpreted as a map, which we take to be sufficiently differentiable, from $\mathbb{R}^l_+$ to $\mathbb{R}^l$, where $\mathbb{R}^l$ is Cartesian l-space and $\mathbb{R}^l_+ = \{p \in \mathbb{R}^l / p_l \geq 0\}$. An economic equilibrium is a set of prices $p = (p_1,\ldots, p_i)$, for which excess demand is zero, that is, $z(p) = 0$.

Economic theory imposes some conditions on the function z which go as follows.

First and foremost is Walras's Law, which is expressed simply by $p{\cdot}z(p) = 0$ (inner product). Written out, this is

$$\sum_{i=1}^{l} p_i \cdot z_i(p_1, \ldots p_l) = 0$$

and states that the value of the excess demand is zero. This is a budget constraint which asserts that the excess demand is consistent with the total assets of the economy. It can be proved from a reasonable microeconomic foundation, as can be seen below.

Second is the homogeneity condition $z(\lambda p) = z(p)$ for all $\lambda > 0$. Changing all prices by the same factor does not affect excess demand. This condition reflects the fact that the economy is self-contained; prices are not based on anything lying outside the model.

The final condition is the boundary condition that $z_i(p) \geq 0$ if $p_i = 0$. This may be interpreted as: if the good is free, then there will be a non-negative excess demand for it.

The following result and its generalizations and ramifications lie at the heart of economic theory.

Existence th

Suppose that an excess demand z satisfies Walras's Law, homogeneity, and the boundary condition. Then there is a price equilibrium.

We will give the proof under the additional mild non-degeneracy condition that the derivative of z is non-singular somewhere on the boundary of $\mathbb{R}^l_+$. This proof is based on Sard's theorem and the inverse function theorem, two basic theorems of global analysis.

Consider a differentiable map f from a set U contained in $\mathbb{R}^k$ to $\mathbb{R}^n$. A vector $y \in \mathbb{R}^n$ is said to be a *regular* value if at each point $x \in U$ with $f(x) = y$, the derivative $Df(x) : \mathbb{R}^k \to \mathbb{R}^n$ is surjective. A subset of $\mathbb{R}^n$ is of *full measure* if its complement has measure zero.

Sard's Theorem

If a map $f : U \to \mathbb{R}^n$ is of class (sufficient differentiability Cr, r $> k - n$, 0) then the set of regular values of f has full measure.

A subset V of $\mathbb{R}^n$ is called a k-dimensional *submanifold* if for each point, there is a neighbourhood U in V and a change of coordinates of $\mathbb{R}^n$ which throws U into a coordinate subspace of dimension k.

Inverse Function Theorem

If $y \in \mathbb{R}^n$ is a regular value of a smooth map $f: U \to \mathbb{R}^n$, $U \subset \mathbb{R}^k$, then either $f^{-1}(y)$ is empty or it is a submanifold of dimension $k - n$.

Let us sketch out the proof of the existence theorem. Define the space of normalized prices by

$$\Delta_1 = \left\{ p \in \mathbb{R}^l_+ / \sum p_l = 1 \right\}.$$

A space auxiliary to the commodity space is defined by

$$\Delta_0 = \left\{ z \in \mathbb{R}^l / \sum z_l = 0 \right\}.$$

From the excess demand map $z : \mathbb{R}^l_+ - 0 \to \mathbb{R}^l$, define an associated map $\varphi : \Delta_1 \to \Delta_0$ by $\varphi(p) = Z(p) - \varphi z^i(p)p$. Note that $\varphi(p)$ is well-defined (i.e. $\varphi(p) \in \Delta_0$) and also smooth. Note moreover that if $\varphi(p) = 0$, then $z(p) = 0$, and that p is a price equilibrium. This follows from Walras's Law as follows. If $\varphi(p) = 0$, then z-$(p) = \Sigma z^i(p)p$ and so $p \cdot z(p) = \Sigma z^i(p)p \cdot p = 0$. Therefore $\Sigma z^i(p) = 0$, since $p \cdot p \neq 0$. By the previous equation $z(p)$ must be zero.

The boundary condition on z implies that φ satisfies a similar boundary condition. That is, if $p_i = 0$ then $\varphi_i(p) = z_i(p) \geq 0$.

It is now sufficient to show that $\varphi(p) = 0$ for some $p \in \Delta_1$. The argument for this proceeds by defining yet another map $\hat{\varphi}$ by

$$\hat{\varphi}(p) = \frac{\varphi(p)}{\|\varphi(p)\|}$$

where $E = \varphi^{-1}(0)$ and S^{1-2} is the set of unit vectors in Δ_0.

By definition the set E is the set of price equilibria, which is to be shown not empty.

Let p_0 be a price vector on the boundary of Δ_1 where the derivative $D\varphi(p_0)$ is non-degenerate (our special hypothesis implies the existence of this p_0). One applies Sard's theorem to obtain'a regular value y of $\hat{\varphi}$ in S^{1-2} near $\hat{\varphi}(p_0)$, where $\hat{\varphi}^{-1}(y)$ is non-empty.

From the inverse function theorem it follows that $\hat{\varphi}^{-1}(y)$ a smooth curve in Δ_1, (a 1-dimensional submanifold). From the boundary condition and a short argument which we omit, it follows that this curve cannot leave Δ_1.

Since the curve $\hat{\varphi}^{-1}(y)$ is a closed set in $\Delta_1 - E$ and has no end points (the inverse function theorem implies that) it must tend to E. In particular, E is not empty and therefore the existence theorem is proved.

The above proof is 'geometrically" constructive in that a curve $\gamma = \hat{\varphi}^{-1}(y)$ constructed which leads to a price equilibrium. This picture can be made analytic by showing that y is a solution of the ordinary differential equation 'Global Newton', $d\varphi/dt = \lambda D\varphi(p)^{-1}\varphi(p)$, where λ is +1 or −1 determined by the sign of the determinant of $D\varphi(p)$. As a consequence the Euler method of approximating the solution of an ordinary differential equation can be used to obtain a discrete algorithm for locating a price equilibrium. By an appropriate choice of steps, ± 1, this discrete algorithm near that equilibrium is Newton's Method; thus the appellation 'Global Newton' for the differential equation.

One would like to understand the process of convergence to equilibrium in terms of decentralized mechanics of price adjustment. Unfortunately the situation in this respect is unclear.

Next we give a brief picture of how global analysis relates to a pure exchange economy. This will allow a microeconomic derivation of the excess demand function discussed above, so that the existence theorem just proved will imply an existence theorem for a price equilibrium of a pure exchange economy. Continuing in this framework one can prove Debreu's theorem on

generic finiteness of price equilibria, by putting the structure of a differentiable manifold on the big set of price equilibria. The equilibrium manifold is a natural setting for comparative statics.

A trader's preferences will be supposed to be represented by a smooth utility function $u : P \to \mathbb{R}$, where $p = \{x \in \mathbb{R}^l, x_i > 0\}$, is commodity space. The indifference surfaces are those $u^{-1}(c) \subset P$. We make strong versions of classical hypotheses on this function.

Monotonicity. The gradient, grad $u(x)$, has positive coordinates.

Convexity. The second derivative D2u(x) is negative definite on the tangent space at x of the corresponding indifference surface.

Boundary condition. The indifference surfaces are closed sets in $\mathbb{R}^l$ (not just P).

From the utility function, one defines for the individual trader a demand function. $f : \mathbb{R}^l_+ \times \mathbb{R}_+ \to P$ of prices $p \in \mathbb{R}^l_+$ and wealth $w > 0$. For this, consider the budget set $B_{p\cdot w} = \{x \in P | \, p \cdot x = w\}$. Then $f(p, w)$ is the maximum of f. on $B_{p.w}$.

One can prove:

Proposition The demand function f satisfies

(a) grad $u(f(p, w) = \lambda p$, for some $\lambda > 0$
(b) $p{\cdot}f(p, w) = w$
(c) $f(\lambda p, \lambda w) = f(p, w)$ any $\lambda > 0$
(d) f is smooth.

A pure exchange economy will be a set of m traders, each with preferences as discussed above, associated to utility functions $u_i \, i = 1, \ldots, m$ defined on the same commodity space P. Also associated to the ith trader is an endowment vector $e_i \in P$. At prices p, this trader's wealth is the value of his endowment $p{\cdot}e_i = w_i$. A *state* is an allocation $(x_1,\ldots,x_m)$, $x_i \in P$ and a price system $p \in \mathbb{R}_+$.

Feasibility is the condition:

$$(F) \sum x_i = \sum e_i$$

A kind of satisfaction condition of a state is

(*S*) For each *i*, x_i *maximizes* u_i *on the budget set*

$$B = \{y \in P | \, p.y = p.e_i\},$$

An economic equilibrium of a pure exchange economy $(e_1,\ldots,e_m, u_1,\ldots,u_m)$ is a state $[x_1,\ldots, x_m), p]$ satisfying ((*F*)) and (*S*).

Theorem There exists a price equilibrium of every pure exchange economy.

The proof goes by applying the previous existence theorem above. Define the excess demand $Z = D{-}S$ as follows:

$$S(p) = \sum e_i, D(p) = \sum f_i(p, p \cdot e_i),$$

where f_i is the above defined demand of the ith trader. One then shows Walras's Law:

$$\begin{aligned} p \cdot Z(p) &= p \cdot D(p) - p \cdot S(p) \\ &= \sum p \cdot f_i(p, p \cdot e_i) - p \cdot \sum e_i = 0 \end{aligned}$$

using (b) of the proposition above.

Use (c) of the proposition to confirm the homogeneity of z. The use of the boundary condition is more technical. But under the rather strong hypotheses, this gives a fairly complete existence proof for a price equilibrium of a pure exchange economy.

This existence proof extends to prove the Arrow–Debreu theorem in the generality of the latter's *Theory of Value*.

See Also

- General Equilibrium
- Mathematics and Economics
- Regular Economies

Bibliography

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

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.

Global Games

Stephen Morris

Abstract

Global games are a class of incomplete information games where small uncertainty about payoffs implies a significant failure of common knowledge. This allows strategic uncertainty to play a crucial and natural role in pinning down equilibrium play. Introduced in the context of two player, two action games by Carlsson and van Damme (Econometrica 61:989–1018, 1993), global games have inspired tractable modelling frameworks that have been used in a wide variety of applications. This article reviews the key ideas that have played a role in theoretical analysis and creating useful applied tools.

Keywords

Comparative statics; Complete information games; Currency crises; Equilibrium selection; Global games; Incomplete information games; Limit dominance; Limit uniqueness; Multiple equilibria; Noise; Noise independent selection; Regime change game; Risk dominance; Robustness herding; Signalling; Strategic complementarities; Sunspot equilibrium; Supermodularity; Threshold strategies

JEL Classifications

C7

Complete information games often have multiple Nash equilibria. Game theorists have long been interested in finding a way of removing or reducing that multiplicity. Carlsson and van Damme (1993) (CvD) introduced an original and attractive approach to doing so. A complete information model entails the implicit assumption that there is common knowledge among the players of the payoffs of the game. In practice, such common knowledge will often be lacking. CvD suggested a convenient and intuitive way of relaxing that common knowledge assumption: suppose that, instead of observing payoffs exactly, payoffs are observed with a small amount of continuous noise; and suppose that – before observing their signals of payoffs – there was an *ex ante* stage where any payoffs were possible. Based on the latter feature, CvD dubbed such games 'global games'. It turns out that there is a unique equilibrium in the game with a small amount of noise. This uniqueness remains no matter how small the noise is and is independent of the distribution of the noise. Since complete information, or common knowledge of payoffs, is surely always an idealization anyway, the play selected in the global game with small noise can be seen as a prediction for play in the underlying complete information game.

The following example illustrates the main idea. There are two players each of whom must decide whether to invest or not invest. Action 'not invest' always gives a payoff of 0. Action 'invest' always gives a payoff of θ; but there are strategic complementarities, and if the other player does not invest, then the player loses 1. Thus the payoff matrix is:

	Invest	Not invest
Invest	θ, θ	θ–1, 0
Not invest	0, θ–1	0, 0

Let us first examine the Nash equlibria of this game when θ is common knowledge. If $\theta < 0$, then 'invest' is a strictly dominated action for each player, and thus 'not invest, not invest' is the unique Nash (and dominant strategies) equilibrium. If $\theta > 1$, then 'not invest' is a strictly dominated action for each player, and 'invest, invest' is the unique Nash (and dominant strategies) equilibrium. The multiplicity case arises if $0 < \theta < 1$. In this case, there are two strict Nash equilibria (both not invest and both invest) and there is also a strictly mixed Nash equilibrium.

But suppose the players do not exactly observe θ. Suppose for convenience that each player believes that θ is uniformly distributed on the real line (thus there is an 'improper' prior with infinite mass: this does not cause any technical or conceptually difficulties as players will always condition on signals that generate 'proper' posteriors).

Suppose that each player observes a signal $x_i = \theta + \sigma\varepsilon_i$, where each ε_i is independently normally distributed with mean 0 and standard deviation 1.

In this game of incomplete information, a pure strategy for player i is a mapping s_i: $\mathbb{R} \to \{\text{Invest, Not invest}\}$. Suppose player 1 was sure that player 2 was going to follow a 'threshold' strategy where she invested only if her signal were above k, so

$$s_2(x_2) = \begin{cases} \text{Invest, if } x_2 > k \\ \text{Not invest, if } x_2 \leq k \end{cases}$$

What is player 1's best response? First, observe that his expectation of θ is x_1. Second, note that (under the uniform prior assumption) his posterior on θ is normal with mean x_1 and variance σ^2, and thus his posterior on x_2 is normal with mean x_1 and variance $2\sigma^2$. Thus his expectation that player j will not invest is $\Phi\left(\frac{1}{\sqrt{2}\sigma}(k - x_1)\right)$, where Φ is cumulative distribution of the standard normal. Thus his expected payoff is

$$x_1 - \Phi\left(\frac{1}{\sqrt{2}\sigma}(k - x_1)\right), \tag{1}$$

and player 1 will invest if and only if (1) is positive. Now if we write $b(k)$ for the unique value of x_1 setting (1) equal to 0 (this is well defined since (1) is strictly increasing in x_1), the best response of player 1 is then to follow a cut-off strategy with threshold equal to $b(k)$. Observe that as $k \to -\infty$ (player 2 always invests), (1) tends to x_1, so $b(k) \to 0$. As $k \to \infty$ (player 2 never invests), (1) tends to $x_1 - 1$, so $b(k) \to 1$. Also observe that if $k = \frac{1}{2}$, then $b(k) = \frac{1}{2}$, since if player 1 observes signal $\frac{1}{2}$, his expectation of θ is $\frac{1}{2}$ but he assigns probability $\frac{1}{2}$ to player 2 not investing. Finally, observe that (by total differentiation)

$$b'(k) = \frac{1}{1 + \dfrac{\sqrt{2\pi}}{\varphi\left(\dfrac{1}{\sqrt{2}\sigma}(k - x_1)\right)}} \in (0, 1),$$

so $b(k)$ is strictly increasing in k and we can immediately conclude that there is a unique 'threshold' equilibrium where each player uses a threshold of $\frac{1}{2}$.

The strategy with threshold $\frac{1}{2}$ is in fact the unique strategy surviving iterated deletion of (interim) strictly dominated strategies. In fact, a strategy s survives n rounds of iterated deletion of strict dominated strategies if and only if

$$s(x) = \begin{cases} \text{Invest, if } x > b^n(1) \\ \text{Not invest, if } x < b^n(0) \end{cases}$$

where $b^n(k) = \overbrace{b(b(\ldots b}^{n \text{ times}}(k))\big)$.

The key intuition for this example is that the uniform prior assumption ensures that each player, whatever his signal, attaches probability $\frac{1}{2}$ to his opponent having a higher signal and probability $\frac{1}{2}$ to him having a lower signal. This property remains true no matter how small the noise is, but breaks discontinuously in the limit: when noise is zero, he attaches probability 1 to his opponent having the same signal.

In this article, I will first report how Carlsson and van Damme's (1993) analysis can be used to give a complete general analysis of two player two action games. I will then report in turn theoretical extensions of their work and a literature that has used insights from global games in economic applications. This dichotomy is somewhat arbitrary (many 'applied' papers have significant theoretical contributions) but convenient.

Two-Player, Two-Action Games

Let the payoffs of a two-player, two-action game be given by the following matrix:

	A	B
A	θ_1, θ_2	θ_3, θ_4
B	θ_5, θ_6	θ_7, θ_8

Thus a vector $\theta \in \mathbb{R}^8$ describes the payoffs of the game and is drawn from some distribution. For a generic choice of θ, there are three possible configurations of Nash equilibria.

1. There is a unique Nash equilibrium with both players using strictly mixed strategies.

2. There is a unique strict Nash equilibrium with both players using pure strategies.
3. There are two pure strategy strict Nash equilibria and one strictly mixed strategy Nash equilibrium.

In the last case, Harsanyi and Selten (1988) proposed the criterion of risk dominance to select among the multiple Nash equilibria. Suppose that (A, A) and (B, B) are strict Nash equilibria of the above game (that is, $\theta_1 > \theta_5$, $\theta_7 > \theta_3$, $\theta_2 > \theta_4$ and $\theta_8 > \theta_6$). Then (A, A) is a risk dominant equilibrium if

$$(\theta_1 - \theta_5)(\theta_2 - \theta_4) > (\theta_7 - \theta_3)(\theta_8 - \theta_6).$$

Generically, exactly one of the two pure Nash equilibria will be risk dominant.

Now consider the following incomplete information game $G(\sigma)$. Each player i observes a signal $x_i = \theta + \sigma\varepsilon_i$, where the ε_i are eight-dimensional noise terms. Thus we have an incomplete information game parameterized by $\sigma \geq 0$. A strategy for a player is a function from possible signals $\mathbb{R}^8$ to the action set $\{A, B\}$. For any given strategy profile of players in the game $G(\sigma)$ and any actual realization of the payoffs θ, we can ask what is the distribution over action profiles in the game (averaging across signal realizations).

Theorem For any sequence of games $G(\sigma^k)$ where $\sigma^k \to 0$ and any sequence of equilibria of those games, average play converges at almost all payoff realizations to the unique Nash equilibrium (if there is one) and to the risk dominant Nash equilibrium (if there are multiple Nash equilibria).

This is shown by the main result of Carlsson and van Damme (1993) in cases (2) and (3) above. They generalize the argument from the example described above to show that, if an action is part of a risk dominant equilibrium or a unique strict Nash equilibrium of the complete information game θ, then – for sufficiently small σ – that action is the unique action surviving iterated deletion strictly dominated strategies. Kajii and Morris (1997) show that, if a game has a unique correlated equilibrium, then that equilibrium is 'robust to incomplete information', that is, will continue to be played in some equilibrium if we change payoffs with small probability. This argument can be extended to show the theorem for case (1) (the extension is discussed in Morris and Shin 2003).

Theoretical Extensions; Many Players and Many Actions

Carlsson and van Damme (1993) dubbed their perturbed games for the two player, two action case 'global games' because all possible payoff profiles were possible. They showed that there was a general way of adding noise to the payoff structure such that, as the noise went to zero, there was a unique action surviving iterated deletion of (interim) dominated strategies (a 'limit uniqueness' result). And they showed that the action that got played in the limit was independent of the distribution of noise added (a 'noise independent selection' result). Their result does not extend in general to many player many action games. In discussing known extensions, we must carefully distinguish which of their results extend.

Frankel et al. (2003) consider games with strategic complementarities (that is, supermodular payoffs). Rather than allow for all possible payoff profiles, they restrict attention to a one-dimensional set of possible payoff functions, or states, which are ordered so that higher states lead to higher actions. The idea of 'global' games is captured by a 'limit dominance' property: for sufficiently low values of θ, each player has a dominant strategy to choose his lowest action, and that for sufficiently high values of θ, each player has a dominant strategy to choose his highest action. Under these restrictions, they are able to present a complete analysis of the case with many players, asymmetric payoffs and many actions. In particular, a limit uniqueness result holds: if each player observes the state with noise, and the size of noise goes to zero, then in the limit there is a unique strategy profile surviving iterated deletion of strictly dominated strategies. Note that while Carlsson and van Damme required no strategic complementarity and other monotonicity properties, when there are multiple equilibria in a two-player,

two-action game – the interesting case for Carlsson and van Damme's analysis – there are automatically strategic complementarities.

Within this class of monotonic global games where limit uniqueness holds, Frankel et al. (2003) also provide sufficient conditions for 'noise independent selection'. That is, for some complete information games, which action gets played in the limit as noise goes to zero does not depend on the shape of the noise. Frankel et al. (2003) show that a generalization of the potential maximizing action profile is sufficient for noise independent selection. This sufficient condition encompasses the risk dominant selection in two player binary action games; the selection of the 'Laplacian' action (a best response to a uniform distribution over others' actions) in many player, binary action games (Morris and Shin 2003). It also yields unique predictions in the continuum player currency crisis of Guimaeres and Morris (2004) and in two-player, three-action games with symmetric payoffs. Morris and Ui (2005) give further sufficient conditions for equilibria to be 'robust to incomplete information' in the sense of Kajii and Morris (1997), which will also ensure noise independent section.

However, Frankel et al. (2003) also provide an example of a two-player, four-action, symmetric payoff game where noise independent selection fails. Thus there is a unique limit as the noise goes to zero, but the nature of the limit depends on the exact distribution of the noise. Carlsson (1989) gave a three-player, two-action example in which noise independent selection failed. Corsetti et al. (2004) describe a global games model of currency crises, where there is a continuum of small traders and a single large trader. This is thus a many-player, two-action game with *asymmetric* payoffs. The equilibrium selected as noise goes to zero depends on the relative informativeness of the large and small traders' signals. This is thus an application where noise-independent selection fails.

More limited results are available on global games without supermodular payoffs. In many applications – such as bank runs – there are some strategic complementarities but payoffs are not supermodular everywhere: conditional on enough people running on the bank to cause collapse, I am better off if I run if few people run and share in the liquidation of the bank's assets. An important paper of Goldstein and Pauzner (2005) has shown equilibrium uniqueness for 'bank run payoffs' – satisfying a single crossing property – with uniform prior and uniform noise. This analysis has been followed in a number of applications. They establish that there is a unique equilibrium in threshold strategies and there are no nonthreshold equilibria. However, their analysis does not address the qst of which strategies survive iterated deletion of strictly dominated strategies. Morris and Shin (2003) discuss how the existence of a unique threshold equilibrium can be established more generally under a signal crossing property on payoffs and a monotone likelihood ratio property on signals (not required for global games analysis with supermodular payoffs); however, these arguments do not rule out the existence of non-monotonic equilibria. Results of van Zandt and Vives (2007) can be used more generally to establish the existence of a unique monotone equilibrium under weaker conditions than supermodularity.

The original analysis of Carlsson and van Damme (2003) relaxed the assumption of common knowledge of payoffs in a particular way: they assumed that there was a common prior on payoffs and that each player observes a small conditionally independent signal of payoffs. This is an intuitively small perturbation of the game and this is the perturbation that has been the focus of study in the global games literature. However, when the noise is small one can show that types in the perturbed game are close to common knowledge types in the product topology on the universal type space: that is, for each type t in the perturbed game, there is a common knowledge type t' such that type t and t' almost agree in their beliefs about payoffs, they almost agree about their beliefs about the opponents' beliefs, and so on up to any finite level. Thus the 'discontinuity' in equilibrium outcomes in global games when noise goes to zero is illustrating the same sensitivity to higher order beliefs of the famous example of Rubinstein (1989). Now we can ask: how general is the phenomenon that Rubinstein

(1989) and Carlsson and van Damme (1993) identified? That is, for which games and actions is it the case that, under common knowledge, the action is part of an equilibrium (and thus survives iterated deletion of strictly dominated strategies) but for a type 'close' to common knowledge of that game, that action is the unique action surviving iterated deletion of strictly dominated strategies. Weinstein and Yildiz (2007) shows that this is true for every action surviving iterated deletion of strictly dominated strategies in the original game. This observation highlights the fact that the selections that arise in standard global games arise not just because one relaxes common knowledge, but because it is relaxed in a particular way: the common prior assumption is maintained and outcomes are analysed under that common prior, and the noisy signal technology ensures particular properties of higher-order beliefs, that is, that each player's beliefs about how other players' beliefs differ from his is not too dependent on the level of his beliefs.

Applications; Public Signals and Dynamic Games

Complete information models are often used in applied economic analysis for tractability: the complete information game payoffs capture the essence of the economic problem. Presumably there is not in fact common knowledge of payoffs, but if asymmetries of information are not the focus of the economic analysis, this assumption seems harmless. But complete information games often have multiple equilibria, and policy analysis – and comparative statics more generally – are hard to carry out in multiple equilibrium models. The global games analysis surveyed above has highlighted how natural relaxations of the common knowledge assumptions often lead to intuitive selections of a unique equilibrium. This suggests these ideas might be useful in applications. Fukao (1994) and Morris and Shin (1995) were two early papers that pursued this agenda. The latter paper – published as Morris and Shin (1998) – was an application to currency crises, where the existing literature builds on a dichotomy between 'fundamentals-driven' models and multiple equilibrium or 'sunspot' equilibria views of currency crises. This dichotomy does not make sense in a global games model of currency crises: currency attacks are 'self-fulfilling' – in the sense that speculators are attacking only because they expect others to do so – but their expectations of others' behaviour may nonetheless be pinned down by higher order beliefs (see Heinemann 2000, for an important correction of the equilibrium characterization in Morris and Shin 1998). Morris and Shin (2000) laid out the methodological case for using global games as a framework for economic applications. Morris and Shin (2003) surveys many early applications to currency crises, bank runs, the design of international institutions and asset pricing, and there have been many more since. Rather than attempt to survey these applications, I will highlight two important methodological issues – public signals and dynamics – that have played an important role in the developing applied literature.

To do this, it is useful to consider an example that has become a workhorse of the applied literature, dubbed the 'regime change' game in a recent paper of Angeletos et al. (2007). The example comes from a 1999 working paper on 'Coordination Risk and the Price of Debt' presented as a plenary talk at the 1999 European meetings of the Econometric Society, eventually published as Morris and Shin (2004). A continuum of players must decide whether to invest or not invest. The cost of investing is c. The payoff to investing is one if the proportion investing is at least $1 - \theta$, 0 otherwise. If there is common knowledge of θ and $6\ \theta \in (0,1)$, there are multiple Nash equilibria of this continuum player complete information game: 'all invest' and 'all not invest'. But now suppose that θ is normally distributed with mean y and standard deviation τ. Each player in the continuum population observes the mean y (which is thus a public signal of θ). But in addition, each player i observes a private signal x_i, where the private signals are distributed in the continuum population with mean θ and standard deviation σ

(that is, as in the example at the beginning of this article). Morris and Shin (2004) show that the resulting game of incomplete information has a unique equilibrium if and only if $\sigma \leq \sqrt{2\pi}\tau^2$, that is, if private signals are sufficiently accurate relative to the accuracy of public signals. This result is intuitive: we know that if there is common knowledge of θ, there are multiple equilibria. A very small value of τ means that the public signal is very accurate and there is 'almost' common knowledge.

This result makes it possible to conduct comparative statics within a unique equilibrium not only in the uniform prior, no 'public' information, limit but also with non-trivial public information. A distinctive comparative static that arises is that the unique equilibrium is very sensitive to the public signal y, even conditioning on the true state θ (see Morris and Shin 2003, 2004; Angeletos and Werning 2006). This is because, for each player, the public signal y becomes a more accurate prediction of others' behaviour than his private signal, even if they are of equal precision.

But the sensitivity of the uniqueness result to public signals also raises a robustness qst. Public information is endogenously generated in economic settings, and thus a qst that arises in many dynamic applications of global games in general and the regime change game in particular is when endogenous information generates enough public information to get back multiplicity (Tarashev 2003; Dasgupta 2007; Angeletos et al. 2006, 2007; Angeletos and Werning 2006; Hellwig et al. 2006). This literature has highlighted the importance of endogenous information revelation and the variety of channels through which such revelation may lead to multiplicity or enhance uniqueness. In addition, these and other dynamic applications of global games raise many other important methodological issues, such as the interaction between the global game uniqueness logic and 'herding' – informational externalities in dynamic settings without payoff complementarities – and 'signalling' – biasing choices from static best responses in order to influence opponents' beliefs in the future.

See Also

- ▶ Coordination Problems and Communication
- ▶ Currency Crises
- ▶ Purification
- ▶ Quantal Response Equilibria

Bibliography

Angeletos, G.-M., C. Hellwig, and A. Pavan. 2006. Signalling in a global game: Coordination and policy traps. *Journal of Political Economy* 114: 452–484.

Angeletos, G.-M., C. Hellwig, and A. Pavan. 2007. Dynamic global games of regime change: Learning, multiplicity and timing of attacks. *Econometrica* 75: 711–756.

Angeletos, G.-M., and I. Werning. 2006. Crises and prices: Information aggregation, multiplicity and volatility. *American Economic Review* 96: 1720–1736.

Carlsson, H. 1989. *Global games and the risk dominance criterion*. Department of Economics/University of Lund. Unpublished manuscript.

Carlsson, H., and E. van Damme. 1993. Global games and equilibrium selection. *Econometrica* 61: 989–1018.

Corsetti, G., A. Dasgupta, H.S. Shin, and S. Morris. 2004. Does one Soros make a difference? The role of a large trader in currency crises. *Review of Economic Studies* 71: 87–114.

Dasgupta, A. 2007. Coordination and delay in global games. *Journal of Economic Theory* 134: 195–225.

Frankel, D., S. Morris, and A. Pauzner. 2003. Equilibrium selection in global games with strategic complementarities. *Journal of Economic Theory* 108: 1–44.

Fukao, K. 1994. Coordination failures under incomplete information and global games. Discussion Paper No. A299. Institute of Economic Research/ Hitotsubashi University.

Goldstein, I., and A. Pauzner. 2005. Demand-deposit contracts and the probability of bank runs. *Journal of Finance* 60: 1293–1327.

Guimaeres, B., and Morris, S. 2004. Risk and wealth in a model of self-fulfilling currency attacks. Discussion Paper No. 1433R. Cowles Foundation.

Harsanyi, J., and R. Selten. 1988. *A general theory of equilibrium selection in games*. Cambridge, MA: MIT Press.

Heinemann, F. 2000. Unique equilibrium in a model of self-fulfilling currency attacks: Comment. *American Economic Review* 90: 316–318.

Hellwig, C., A. Mukherji, and A. Tsyvinski. 2006. Self-fulfilling currency crises: The role of interest rates. *American Economic Review* 96: 1769–1787.

Kajii, A., and S. Morris. 1997. The robustness of equilibria to incomplete information. *Econometrica* 65: 1283–1309.

Morris, S., and Shin, H.S. 1995. Informational events that trigger currency attacks. Working Paper No. 95-24. Federal Reserve Bank of Philadelphia.
Morris, S., and H.S. Shin. 1998. Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review* 88: 587–597.
Morris, S., and H.S. Shin. 2000. Rethinking multiple equilibria in macroeconomics. In *NBER Macroeconomics Annual 2000*, ed. B.S. Bernanke and K. Rogoff. Cambridge, MA: MIT Press.
Morris, S., and H.S. Shin. 2003. Global games: Theory and applications. In *Advances in economics and econometrics (Proceedings of the eighth world congress of the econometric society)*, ed. M. Dewatripont, L. Hansen, and S. Turnovsky. Cambridge: Cambridge University Press.
Morris, S., and H.S. Shin. 2004. Coordination risk and the price of debt. *European Economic Review* 48: 133–153.
Morris, S., and T. Ui. 2005. Generalized potentials and robust sets of equilibria. *Journal of Economic Theory* 124: 45–78.
Rubinstein, A. 1989. The electronic mail game: Strategic behavior under almost common knowledge. *American Economic Review* 79: 385–391.
Tarashev, N. 2003. Speculative attacks and informational role of the interest rate. *Journal of the European Economic Association* 5: 1–36.
van Zandt, T., and X. Vives. 2007. Monotone equilibria in Bayesian games of strategic complementarities. *Journal of Economic Theory* 134: 339–360.
Weinstein, J., and M. Yildiz. 2007. A structure theorem for rationalizability with application to robust predictions of refinements. *Econometrica* 75: 365–400.

Globalization

William Easterly

Abstract

The passion that surrounds the vague term 'globalization' is best seen as a proxy for the long-standing debate about free-market capitalism. The zero-sum mindset, the difference between Pareto superiority and common norms of fairness, and the belief that all outcomes are caused by an intentional agent often cause communication problems between non-economists and free-market economists, who themselves often exaggerate what 'free-market reforms' can accomplish and endorse overly ambitious programmes of change ('shock therapy'), underestimating problems of transition and the second best. Economists could try to understand the protests against 'globalization' rather than dismissing them out of hand.

Keywords

Anti-capitalism; Asian miracle; Banking crises; Business networks; Calhoun, J. C.; Carlyle, T.; Contract enforcement; Creative destruction; Development economics; Dismal science; Economic growth; Financial liberalization; Financial regulation; Gains from trade; Globalization; Inequality; International trade; Invisible hand; Lenin, V. I.; Outsourcing; Poverty; Poverty alleviation; Reform consultants; Second best; Shock therapy; Slavery; Spontaneous order; Structural adjustment; Stylized facts; Total factor productivity; Washington Consensus

JEL Classifications

O5

'Globalization' is a word that gets both its proponents and opponents very agitated. But what exactly is it? What is the globalization debate really about?

The answer is that the globalization debate is about a surprisingly large number of issues, including some that lie outside of economics. A non-exhaustive list of issues derived from a reading of the writings of both economists and non-economists (see a very partial list of references in the bibliography) follows:

1. Liberalization versus regulation of international trade, capital movements, and migration.
2. Market imperfections that arise with (either domestic or international) goods markets, capital markets, privatization, macroeconomic crises, intellectual property rights, and so on.
3. Evaluation of the performance of the International Monetary Fund (IMF) and the World

Bank, including in particular their policy prescriptions (the 'Washington Consensus', 'shock therapy', or 'structural adjustment').
4. Effects of freer trade and capital movements on rich country workers ('outsourcing') and on poor country workers ('sweatshops').
5. Extreme world inequality and poverty.
6. Capitalism ('neoliberalism') versus alternative systems.
7. Westernization/Americanization versus local culture.
8. Unequal distribution of political power between the West (both Western governments and corporations) and the Rest.
9. Effect of global economic growth on the environment.
10. Western imperialism and military intervention in the rest of the world.

Arguably, the vagueness of a term that includes at least ten separate debates has done a disservice to economic and political debate, causing many 'globalization' debate participants to think they disagree with people with whom they really agree, or to think they agree with people with whom they really disagree. It also explains some of the difficulties in communication between economists and non-economists about globalization, because the two groups really have different debates in mind. Economists (including those identified as 'globalization critics') have focused largely on issues 1–5, while the non-economists – though not ignoring 1–5 – seem to have something else in mind like 6–10.

For example, Dani Rodrik (1997) and Joseph Stiglitz (2002), who have both acquired a reputation as globalization critics by focusing mainly on issues 1–3, are embraced eagerly by some 'globalization protesters' whose main issue is really 6: the critique of capitalism (sometimes called 'neoliberalism'). This is not meant as a criticism of Rodrik and Stiglitz; rather, it highlights the confusion that exists when two prominent mainstream economists who are talking about tinkering with and fine- tuning capitalist markets are seen as allies by those who are opposed to free market capitalism.

This article can hardly do justice to the complexity of all of these debates, nor is there much hope of getting everyone to discontinue the almost criminally vague use of the term 'globalization' in debate. The article argues that most of the energy in the debate indeed comes from the clash of attitudes – enthusiastic and antipathetic – towards capitalism and free markets.

This article thus focuses on two key themes about the globalization debate. First, I give some intellectual history of the debate about capitalism (issue 6), which will place in perspective some of today's globalization debate including that by the non-economists. This has the objective of dispelling some of the puzzlement that many economists feel about the sound and fury surrounding globalization, through realization that it is partly just another manifestation of a long intellectual debate about capitalist free markets, which economists have been engaged in for decades if not centuries. Second, the article tries to place the antipathy towards free markets in contemporary perspective by discussing whether overly simplistic models and unrealistic promises of quick and sizeable results from 'globalization' for poor countries have further fuelled this antipathy. I consider at the same time whether the zeal of the globalizers may have led them to endorse counterproductive and unrealistic attempts at wholesale social transformation, which generate an even more severe backlash.

Let's start with the long-standing debate about capitalism. Intellectual history makes clear how the gains from trade (in goods, finance, and labour services) under capitalism amount to such a revolutionary idea that economists are often its lone proponents in the wilderness. There are three major habits of thinking that create difficulty in communication between economists and non-economists on gains from trade. One is the mindset that holds that economic interactions are zero-sum games (a partially understandable mindset when capitalists have such skeletons in the closet as military conquest, colonization, slavery, predatory behaviour by firms, and so on). The second is the difference between economists' notion of Pareto-superior outcomes and common social norms of fairness. The third barrier to

communication is the difficulty of accepting the economists' notion of the invisible hand that creates spontaneous outcomes not designed by anyone, where the common habit of thinking is that a good or bad outcome must be the result of intentional action by a good or bad agent.

To start with the zero-sum mindset, one early father of Christianity, St. Jerome, thought any wealth was automatically 'unjust riches', since 'no one can possess them except by the loss and ruin of others'. St. Augustine put it more tersely: 'If one does not lose, the other does not gain' (quoted by Muller 2002, p. 6).

Centuries later, even after Adam Smith and the Industrial Revolution, both sides of the political spectrum still often thought in zero-sum terms. Friedrich Engels wrote that 'the consequences of the factory system' were 'oppression and toil for the many, riches and wealth for the few' (quoted by Muller 2002, p. 180). Henry Adams saw capitalism as a system that divided humanity 'into two classes, one which steals, the other which is stolen from' (quoted by Herman 1997, p. 160).

We are so used to thinking of conservatives as pro-market that it surprises us that some on the Right in the 19th century also attacked free market economics (see Levy 2001, for a fine narrative). The Right's attack on the laissez-faire Left (how things have changed!) was that the latter were hypocritical advocating both capitalism and the end of slavery, because capitalism made 'free' workers no better than slaves. For example, Thomas Carlyle (the man who disliked economists so much that he coined the phrase 'dismal science') told workers: 'you are fallen captive to greedy sons of profit-and-loss; to bad and ever to worse . . . Algiers, Brazil or Dahomey hold nothing in them so authentically *slave* as you are' (Carlyle 1850). This is zero-sum thinking in the extreme!

Similarly, on the American 19th-century Right, John C. Calhoun defended American slavery in 1828 by claiming that industrial capitalism in the North was no better; it caused wages to 'sink more rapidly than the prices of the necessaries of life, till the operatives . . . portion of the products of their labor . . . will be barely sufficient to preserve existence' (quoted by Muller 2002, p. 177). Ironically the great African-American intellectual W. E. B. Du Bois, reached similar conclusions to Calhoun's about industrial capitalism, as he observed it in the South after the Civil War:

> [The] men who have come to take charge of the industrial exploitation of the New South. . .thrifty and avaricious Yankees . . ..For the laborers as such, there is in these new captains of industry neither love nor hate, neither sympathy nor romance; it is a cold question of dollars and dividends. Under such a system all labor is bound to suffer. . . . The results among them, even, are long hours of toil, low wages, child labor, and lack of protection against usury and cheating. (Du Bois 1903)

Lenin famously linked zero-sum Western imperialism and non-zero-sum trade and capital flows. Profits for the companies that follow in the wake of the imperialists are high in the 'backward countries', where the capitalists relocate their capital because 'the price of land is relatively low, wages are low, raw materials are cheap' (Lenin 1917). Lenin may have been the first 20th-century critic of outsourcing.

Today, we see similar zero-sum thinking in globalization critics on the Left and the Right. Oxfam GB (2004, p. 12) identifies such products as Olympic sportswear as forcing labourers into 'working ever-faster for ever-longer periods of time under arduous conditions for poverty-level wages, to produce more goods and more profit' (Statements like this come from an organization that is actually much friendlier to free trade than most non-governmental organizations.)

Global Policy Forum, a popular globalization website elaborates: 'trade is inherently unequal and poor countries seldom experience rising well-being but increasing unemployment, poverty, and income inequality.' Former Tanzanian President Julius Nyerere summarized the zero-sum mindset back in a 1975 state visit to Britain: 'I am poor because you are rich' (quoted in Lindsey 2001, p. 105).

On the Right, there is still today concern about free markets creating winners at the expense of losers. Patrick Buchanan claimed in a 1998 book that free trade causes 'broken homes, uprooted families, vanished dreams, delinquency, vandalism, crime' (quoted in Micklethwait and Wooldridge 2000, p. 282). Edward Luttwak

claimed that global capitalism requires 'harsh laws, savage sentencing, and mass imprisonment' to deal with 'disaffected losers' (1999, p. 236). Although of course the Right in general is today more sympathetic to free market capitalism than the Left, the persistence of this thinking shows how the zero-sum mindset is an independent force from political ideology.

Of course, capitalism/globalization does create losers as well as winners, unleashing gales of creative destruction. Since losers tend to be more vocal than winners, it is easy to understand the perception that the losers outnumber the winners, which then reinforces the already ingrained habit of thinking in zero-sum terms. To complicate matters further, some poorly conceived attempts at rapid transition from non-capitalism to capitalism (for example, 'shock therapy,' to be discussed below) can in fact create more losers than winners.

The second source of communication breakdown about globalization is the difference between economists' general enthusiasm for Pareto improvements and common norms of fairness (see Aisbett 2005, for a provocative discussion). Following Aisbett, let's say for example that a multinational firm opens a factory in a low-income country. Suppose that the new investment enables the firm to double its profits and the newly employed workers in the factory to double their previous incomes. Suppose the workers were formerly part of the extreme poor (conventionally measured as an income of a dollar a day), so that now they have escaped extreme poverty. Who can argue with such a Pareto improvement?

From another perspective, however, what is happened is that a very poor person has gained a dollar (in a 'sweatshop') while a captain of industry previously making, say, $1,000 a day has gained another $1,000. It violates many norms of fairness (abundantly confirmed in the laboratory by experimental economics), when someone already far better off gains 1,000 times more than the less fortunate person from this transaction. To point this out doesn't lead to any obvious conclusions – most economists will say the transaction is still worth doing to relieve absolute poverty, while critics will protest that a more fair division of the gains should be possible (but even if the worker gets a fivefold increase in income – an amazing escape velocity from poverty – while the capitalist just doubles his income, the capitalist's gains are still 250 times larger).

The third barrier to constructive communication about globalization is the common assumption that an outcome must result from an intentional action by an identifiable agent. This couldn't contrast more with the economists' notion of the invisible hand. The intentionality mindset is that 'globalization' represents someone's agenda, and it is to blame for the tragedies of world poverty. To give an illustrative example of this kind of anti-globalization rhetoric (not necessarily representative): the 'transnational corporations . . .expand, invest and grow, concentrating ever more wealth in a limited number of hands. They work in coalition to influence local, national and international institutions'. 'Corporate elites. . .forge common agendas outside the formal institutions of democracy.' They use forums such as the Trilateral Commission, the International Chamber of Commerce, the World Economic Forum, trade associations, and the many national and international business and industrial roundtables' (IFG 2002, p. 140). The participants at such 'posh gatherings. . .chart the course of corporate globalization in the name of private profits . . .' (IFG 2002, p. 4). Searching for whom to blame for the miseries of the poor, the rich multinational corporations make for natural villains (alternative villains are the IMF and World Bank). These are not only villains, but foreign villains! This mindset is further strengthened when corporations (who of course really are self-interested profit-seekers) get caught doing something like despoiling the environment in a poor country or doing shady deals with the local kleptocrats. The economists' idea of a spontaneous system of myriads of uncoordinated agents, with nobody in charge, generating outcomes that are not intended (or even forseeable) by anyone is a lot harder sell.

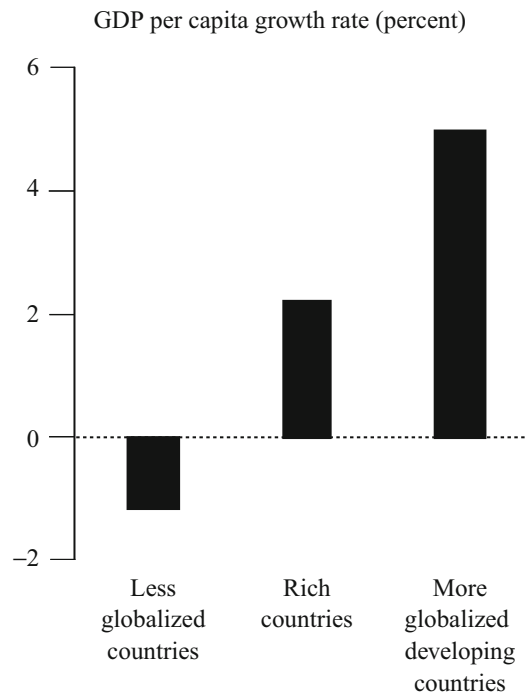
With such fundamental differences in thinking, perhaps we can understand why there is little prospect of a constructive conversation between advocates and opponents of globalization/free

market capitalism/neoliberalism. The World Social Forum, the counterpoint annual meeting to that of the capitalist globalizers at the World Economic Forum in Davos, says in its charter that it is 'an open meeting place for reflective thinking, democratic debate of ideas, free exchange of experiences', except that the debate is limited to 'groups and movements of civil society that are opposed to neoliberalism'. A similar spirit seems to inform the complaint that the case for capitalism arises from 'rationalist constructions of knowledge' featured in such reunions of the 'global managerial class' as 'AEA conventions' (Global Policy Forum 2006). Of course, mainstream economists probably do not seem to their critics much more open to debate on 'neoliberalism'!

Things are made even worse by the second major theme of this article, the overselling of globalization. Simplistic models and promises of quick and sizeable results create expectations, and when these expectations are disappointed (even when the results are gradually and increasingly positive), there is a backlash against globalization.

A classic example of the overselling of globalization is the World Bank (2002) report *Globalization, Growth, and Poverty.* The following graph (Fig. 1), the first one shown in the report, is prominently displayed in the overview (2002, p. 5):

Although never explicitly stated and some caveats are expressed, the impression left with many readers is that being more globalized makes the difference between five per cent per capita growth and minus one per cent per capita growth, which is an amazingly strong claim for the effects of globalization. World Bank researchers reinforce this kind of claim with statements promising that world poverty can be cut in half with policy reforms: 'Poverty reduction – in the world or in a particular region or country – depends primarily on the quality of economic policy. Where we find in the developing world good environments for households and firms to save and invest, we generally observe poverty reduction' (Collier and Dollar 2001). (I have to admit with some embarrassment that this statement was based on one of my own unpublished growth regressions, which eventually showed up in published form in Easterly 2001 making the opposite point – that the growth response to policy reform was disappointing. Regressions can be dangerous!).



Globalization, Fig. 1 Divergent paths of developing countries in the 1990s

The IMF likewise has a standard set of policies that it advocates (together, the IMF's and the World Bank's notion of 'good policies' form what is often called the 'Washington Consensus'), many of which are oriented towards creating freer markets (more 'globalization'). The IMF also claimed that 'Where [good] policies have been sustained, they have raised growth and reduced poverty' (2000).

This is speculation, but some of the World Bank/IMF belief in policy reforms to explain good outcomes may ironically stem from the same intentionality impulse that makes critics blame the World Bank and IMF for bad outcomes. People find it more comfortable to attribute success to the action of a few heroic policy reformers or technocrats (or strong leaders implementing good policies, like Singapore's Lee Kuan Yew),

rather than to some more mysterious bottom-up process of many spontaneous individual entrepreneurs.

Unfortunately, there is little evidence for strong growth effects of policy changes that involve anything less than getting rid of self-destructive extremes (like moving from autarchy to allow some trade, or from hyperinflation to moderate inflation), and even then hardly six percentage points of permanent change in growth, as documented in Easterly (2005). Contrary to the impression conveyed in the foregoing statements, the economics profession actually knows very little about how to raise economic growth over the short to medium run with policy changes in the range in which most countries are operating (see for example the survey in Kenny and Williams 2001). (Besides this, the methodological problem is that countries that are more or less globalized are not defined in terms of policies that promote free trade, free capital movements, free migration, or some other policy measure that features in the debates on globalization. The 'more globalized' countries are defined in terms of outcomes: it is those that are in the top third of countries in terms of the increase in their trade-to-GDP ratios. Defining globalization in terms of one endogenous measure of success that is likely related to other endogenous measures of success – like the GDP growth rates being explained – is rather unfortunate.) Growth in developing countries is extremely volatile (on average 75 per cent of a country's deviation from the global mean per capita growth in a five- or ten-year period disappears in the following period, as pointed out in Easterly et al. 1993, since replicated with more recent data.) Overeager growth-watchers are too quick to proclaim 'growth miracles' and the lessons that allegedly follow from them. As Dixit (2006, p. 23) says,

> At any time, some country is doing well, and academic as well as practical observers are tempted to generalize from its choices and recommend the same to all countries. After a decade or two, this country ceases to do so well, some other country using some other policies starts to do well, and becomes the new star that all countries are supposed to follow.

The success of China and the earlier successes of the East Asian miracles (all associated with great success in global markets) are often used by promoters of globalization to bolster their case. Unfortunately, the implicit promise that such unusually rapid growth (on the order of five per cent per capita) is available to all 'globalizers' rests on very shaky ground. First, such rapid growth is very rare – 1.7 per cent of countries registered five per cent per capita growth or more over 1950–2001, and only 0.7 percent of all half-century country per capita growth episodes since 1820 surpassed five per cent (Maddison 2003). Most rich countries today (usually agreed to have been globalized and capitalist for quite some time) got to be rich by registering something on the order of two per cent per capita growth for one or two centuries.

Things are made worse when casual empiricism is married to simplistic theory. In the simplest textbook model, freedom of trade and capital movements each promotes poverty alleviation, that is, rapid catching up of wages or incomes in capital-scarce, labour-abundant poor countries to wages or incomes in rich countries. According to the model, free trade allows unskilled wages in poor countries to rise rapidly through labour-intensive exports, while free capital movements allow high investment in poor countries to remedy the gap in capital per worker between rich and poor countries. A side effect would be that inequality within the poor country (driven mainly by the differences between labour and capital earnings) should decrease. Even aside from the fact that a vast trade literature does not support most predictions of the first story, the growth and development literature has pointed out that total factor productivity differences between countries are a much more plausible explanation of income differences between countries than differences in capital per worker (Hall and Jones 1999; Klenow and Rodriguez-Clare 1997; Easterly and Levine 2001; Hsieh 2002). Stylized facts on trade, inequality, and poverty do not support the predictions of the simple textbook model where income differences are due to differences in capital per worker. (See Easterly 2006, for a more extensive discussion of these points.)

These false expectations of very rapid growth through globalization (or 'free markets' in general) have arguably done a lot of damage, creating fertile ground for an anti-capitalist backlash in places as diverse as Thailand, South Africa, Russia, Bolivia, Venezuela, Peru, Argentina, Ecuador and Mexico. The critics of globalization can all the more easily seize upon any growth setbacks (such as the Mexico crisis of 1994/95, the East Asian crisis of 1997/98, or the Argentine crisis of 2001, or disappointing growth in Latin America in general since market liberalization in the 1980s), whatever their cause (usually hard to explain anyway in the volatile pattern described earlier), to say 'see, globalization/neoliberalism doesn't work'.

The backlash has been made all the worse because of the overconfidence of IMF and World Bank policymakers (and freelance 'reform consultants') 'globalizing' whole societies that start out with many different barriers to efficient free markets. The economics profession can demonstrate fairly convincingly that some long-run market-friendly policies and institutions are most conducive to prosperity, but it knows very little about the sequencing and the transitional paths of reforms to get from initial conditions to that ideal state. (Lipsey and Lancaster's 1956–57, theory of the second best recognized this problem, but it seems that each new generation must discover it afresh.) It is obvious that different kinds of reforms are complementary to each other – for example, financial market liberalization works well only if there is sufficient transparency of banks to depositors, and a good regulatory and supervisory framework to ensure that banks don't cheat (Barth et al. 2006). Otherwise, financial liberalization often leads to bad loans, enrichment of insiders, and subsequent banking system crises, as abundant experience has already demonstrated.

Yet the usual answer to policy complementarity – 'do everything at once' 'structural adjustment,' or 'shock therapy' – doesn't really escape the curse of the second-best. Policymakers neither know what 'everything' is nor have the ability to change 'everything' at once (or any time soon). The choice is really between large- scale partial reforms (which shock therapy mislabels 'comprehensive reform') and small-scale partial reforms. Any economy is a complex system of informal networks, social norms, relationships, trades, and formal institutions, many of which lie outside the control of the policymaker. As Dixit (2004) points out, an existing network under the current system of rules can at least enforce contracts in that it can threaten to expel any member who cheats another member. Drawing up a brand-new set of rules overnight (like moving abruptly from an interventionist economy to a free-market economy) can have perverse impacts in the short run. It can mean that people can choose to exit the old network (cheating their old partners) because they now have the option of operating under the new system of rules and the new networks generated by the new rules. The net effect can be to disrupt the functioning of the old economy much more than it facilitates the creation of the new economy. This is theoretical speculation at this point, but it does illustrate the potential pitfalls of promising rapid results from rapid reforms. To think that economists could re-engineer the whole society and economy looks in retrospect like the worst kind of intellectual hubris (see McMillan 2007, for a great discussion). Attempts at rapid 'comprehensive' reform of poor countries look a lot like what Karl Popper long ago decried as 'utopian social engineering' versus what has worked for most rich countries to attain prosperity: 'piecemeal democratic reform'.

The most notorious case of this hubris was the attempt to reform the former Soviet Union with 'shock therapy'. Murrell (1992, 1993) – a longtime scholar of centrally planned economies – argued against shock therapy as utopian social engineering. His objections are all the more compelling because they were *ex ante* rather than *ex post*. History vindicated his scathing description of shock therapy at the time:

> There is complete disdain for all that exists . . . History, society, and the economics of present institutions are all minor issues in choosing a reform program. . . Establishment of a market economy is seen as mostly involving destruction. . . shock therapists assume that technocratic solutions are fairly easy to implement. . . One must reject all existing arrangements . . . (Murrell 1993)

Murrell was quick to realize the relevance of Popper for what was later half- jokingly called a Leninist push for free markets in Russia. His quote from Popper in 1992 is a perfect prediction of how Russian reform would fail: 'It is not reasonable to assume that a complete reconstruction of our social system would lead at once to a workable system' (quoted in Murrell 1992). After the former Soviet republics experienced some of the greatest depressions in economic history, the prescience of such viewpoints became apparent.

For its part, IMF- and World Bank-supported 'structural adjustment' was also uncomfortably like 'utopian social engineering', and produced a similar debacle in Africa and Latin America. The resulting anti-market/anti-globalization backlash (in the former Soviet Union as well as Africa and Latin America) was all the more severe because the reforms involved some IMF/World Bank coercion through conditional loans. One can hardly think of a better formula for an anti-capitalist backlash in poor countries than to introduce overambitious, oversold programmes of large-scale 'globalization' reforms imposed by foreigners!

In conclusion, economists are unlikely to find the term 'globalization' a precise enough concept to advance most research agendas. Instead, it mainly seems to point to the long-standing debate about economists' traditional embrace of free-market capitalism. Perhaps some progress in these debates can be made by understanding some of the traditional mindsets that make a system of spontaneous gains from trade such a revolutionary concept. It also would help a great deal if policymakers and the economists advising them did not pursue overambitious attempts at rapid wholesale transformation of the economy and society, and did not exaggerate the likely size and speed of the gains for the economy from such programmes.

Articulate arguments of the case for capitalism/globalization continue to be made in books such as Lindsey (2001) and Wolf (2004). Mishkin (2006) has recently made a fascinating case for the kind of globalization that opponents find most frightening (and even many economists shy from), financial globalization. Despite such eloquent statements, the discomforts caused by the spectre of globalization are unlikely to abate any time soon. Economists can arguably contribute more to the debate by seeking to understand the discomfort rather than dismissing it out of hand.

Bibliography

Aisbett, E. 2005. *Why are the critics so convinced that globalization is bad for the poor?*, Working paper, no. 11066. Cambridge, MA: NBER.

Barth, J.R., G. Caprio, and R. Levine. 2006. *Rethinking bank regulation: Till angels govern*. Cambridge: Cambridge University Press.

Carlyle, T. 1850. The present time. In *Latter-day pamphlets.* London: Chapman & Hall.

Collier, P., and D. Dollar. 2001. Can the world cut poverty in half? How policy reform and international aid can meet international development goals. *World Development* 29: 1767–1802.

Dixit, A.K. 2004. *Lawlessness and economics: Alternative modes of governance*. Princeton: Princeton University Press.

Dixit, A.K. 2006. *Evaluating recipes for development success*, Policy research working paper, no. 3859. Washington, DC: World Bank.

Du Bois, W.E.B. 1903. Of the sons of master and man. Ch. 9. In *The souls of Black folk.* New York: Bartleby.com, 1999.

Easterly, W. 2001. The lost decades: Explaining developing countries' stagnation in spite of policy reform 1980–1998. *Journal of Economic Growth* 6: 135–157.

Easterly, W. 2005. National policies and economic growth: A reappraisal. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North Holland.

Easterly, W. 2006. Globalization, poverty, and all that: Factor endowment versus productivity views. In *Globalization and poverty*, ed. A. Harrison. Chicago: University of Chicago Press.

Easterly, W., and R. Levine. 2001. It's not factor accumulation: Stylized facts and growth models. *World Bank Economic Review* 15(2): 177–219.

Easterly, W., M. Kremer, L. Pritchett, and L. Summers. 1993. Good policy or good luck? Country growth performance and temporary shocks. *Journal of Monetary Economics* 32: 459–483.

Global Policy Forum. 2006. Political struggle will determine better globalization. 15 March. Online. Available at http://www.globalpolicy.org/globaliz/define/2006/03scholte.htm. Accessed 2 Jan 2007.

Global Policy Forum. 2007. International trade and development. Online. Available at http://www.globalpolicy.org/socecon/trade/index.htm. Accessed 2 Jan 2007.

Hall, R.E., and C.L. Jones. 1999. Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* 114: 83–116.
Herman, A. 1997. *The idea of decline in Western history.* New York: Free Press.
Hsieh, C.T. 2002. What explains the industrial revolution in East Asia? Evidence from the factor markets. *American Economic Review* 92: 502–526.
IFG (International Forum on Globalization). 2002. *Alternatives to economic globalization: A better world is possible*. San Francisco: Berrett-Koehler Publishers.
IMF (International Monetary Fund). 2000. *Policies for faster growth and poverty reduction in sub-Saharan Africa and the role of the IMF,* Issues brief. Washington, DC.
Kenny, C., and D. Williams. 2001. What do we know about economic growth? Or, why don't we know very much? *World Development* 29(1): 1–22.
Klenow, P., and A. Rodriguez-Clare. 1997. The neoclassical revival in growth economics: Has it gone too far? *NBER Macroeconomics Annual 1997* 12: 73–103.
Lenin, V.I. 1917. Imperialism, the highest stage of capitalism. Online. Available at http://www.marxists.org/archive/lenin/works/1916/imp-hsc/ch04.htm. Accessed 2 Jan 2007.
Levy, D.M. 2001. *How the dismal science got its name: Classical economics and the ur-text of racial politics*. Ann Arbor: University of Michigan Press.
Lindsey, B. 2001. *Against the dead hand: The uncertain struggle for global capitalism*. New York: Wiley.
Lipsey, R.G., and K. Lancaster. 1956–1957. The general theory of second best. *Review of Economic Studies* 24(1): 11–32.
Luttwak, E. 1999. *Turbo-capitalism: Winners and losers in the global economy.* New York: HarperCollins.
Maddison, A. 2003. *The world economy: Historical statistics*. Paris: OECD.
McMillan, J. 2007. Avoid hubris. In *Reinventing foreign aid*, ed. W. Easterly. Boston: MIT Press.
Micklethwait, J., and A. Wooldridge. 2000. *A future perfect: The challenge and hidden promise of globalization*. New York: Crown Business.
Mishkin, F.S. 2006. *The next great globalization: How disadvantaged nations can harness their financial systems to get rich*. Princeton: Princeton University Press.
Muller, J.Z. 2002. *The mind and the market: Capitalism in modern European thought*. New York: Alfred A. Knopf.
Murrell, P. 1992. Conservative political philosophy and the strategy of economic transition. *East European Politics and Societies* 6(1): 3–16.
Murrell, P. 1993. What is shock therapy? What did it do in Poland and Russia? *Post-Soviet Affairs* 9(2): 111–140.
Oxfam, G.B. 2004. *Play fair at the olympics*. Oxford: Oxfam BG.
Rodrik, D. 1997. *Has globalization gone too far?* Washington, DC: Institute for International Economics.
Stiglitz, J. 2002. *Globalization and its discontents*. New York: Norton.
Wolf, M. 2004. *Why globalization works*. New Haven: Yale University Press.
World Bank. 2002. *Globalization, growth, and poverty: A policy research report*. Washington, DC: World Bank.
World Social Forum. 2007. Charter. Online. Available at http://www.portoalegre2002.org/default.asp. Accessed 30 Jan 2007.

Globalization and Labour

Richard B. Freeman

Abstract

The entry of China, India, and the ex-Soviet countries into the world trading system in the 1990s has made globalization an increasingly important driver of labour outcomes across the world. Through trade, capital flows, the spread of technology and education, the world has begun to move towards a truly global labour market. Still, the dispersion of wages for similar work across countries remains high and immigration is the least developed part of globalization, leaving considerable scope for national labour markets, policies, and institutions to affect wages and worker well-being into the foreseeable future.

Keywords

Brain drain; Child labour; Comparative advantage; Cost of capital; Diffusion of technology; Factor endowments; Factor mobility; Factor price equalization; Fair trade; First-mover advantage; Foreign direct investment; Foreign portfolio investment; Globalization; Globalization and labour; Heckscher–Ohlin trade theory; Higher education; Inequality (global); International migration; International trade; Labour standards; North–South economic relations; Occupational health and safety; Price dispersion; Product life cycle; Production possibility frontier; Purchasing power parity; Ricardian trade theory; Transfer of technology; Wage dispersion

JEL Classifications
F00; J00

Globalization – the export and import of goods and services, international capital mobility, labour mobility, and technical knowledge across national borders – connects economies and influences the economic well-being of workers worldwide. Imports reduce the demand for workers in a country by substituting foreign labour, whose work is embodied in the imports, for domestic labour. Exports increase the demand for workers by selling what workers produce to other countries. Since the wages paid for labour differ among countries, firms have sizable incentives to offshore some jobs to foreign countries, including many service sector jobs that have been historically non-tradable. Capital mobility changes the capital stock with which workers operate, raising or lowering demand for labour. Immigration, business trips, international study and tourism affect the supply and demand for labour. And, most important of all, the flow of knowledge across borders allows countries to improve their technical and economic prowess and operate along the global production possibility frontier even when they lack the scientific base to expand the frontier. Although economic analyses generally treat trade in goods and services, capital flows, labour flows, and the transfer of knowledge separately, these four facets of globalization have feedbacks and connections that help determine their impact on the economy and on the work force.

At the end of the 20th century globalization became a more powerful driver of labour market outcomes than ever before. The collapse of Soviet Communism, China's shift to market capitalism and India's market reforms and entry into the global trading system produced a single economic world based on capitalism and markets. Before those changes, the global economy encompassed roughly half of the world's population – the advanced countries, Latin America, the Caribbean, Africa, and some parts of Asia – while the other half lived in separate economic spaces. Workers in the United States and other higher income countries and in market-oriented developing countries did not face competition from low-wage Chinese or Indian workers nor from workers in the Soviet empire. The entry of these economies into the world trading system in the 1990s increased the global labour pool from approximately 1.46 billion workers to 2.93 billion workers – 'the great doubling' (see Freeman, 2005a, b).

As documented in Freeman (2006, pp. 150–1) and data given on the International Monetary Fund (IMF) website, all aspects of globalization grew at the turn of the 21st century. World trade increased relative to world GDP so that world exports rose to 27 per cent of world GDP in 2005 compared to just 12 per cent of world GDP in 1970. Foreign direct investment, which had been 2–3 per cent of global gross capital formation in the 1970s rose to 7–20 per cent of gross capital formation in the 1990s–2000s. The share of foreign equities in investors' equity portfolios rose from negligible numbers to about 15 per cent in the early 2000s. Immigration from developing countries to advanced countries increased so that in 2000 8.7 per cent of the population in the high income countries had been born elsewhere. The single biggest recipient of immigrants was the United States, where the share of immigrants nearly tripled from 1970 to 2005 and where roughly one in five workers aged 25–39 was foreign-born. As for the transfer of knowledge, university enrolments grew rapidly worldwide and multinationals moved production to developing countries. China, in particular, made rapid gains in measures of technological prowess. According to the Georgia Technology Policy and Assessment Center, between 1993 and 2005, China more than tripled its rating in technological standing (Porter et al. 2006, Table 3).

A comparison of the different facets of globalization indicates that the ratio of immigrants to the world work force is lower than the ratio of trade to goods production and international capital flows to activity in capital markets, which suggests that immigration is the least developed part of globalization. To some extent, this may reflect the greater personal cost in moving from one country to another than to ship goods or capital across

borders. But there is a political economy reason as well. Even countries committed to freer trade and capital mobility do not allow for free immigration.

Economists use two types of models to analyse the effects of globalization on economic performance and the well-being of workers. They use the basic Heckscher–Ohlin model of comparative advantage to analyse trade between advanced countries and developing countries. This model takes country factor endowments (labour skills, natural resources, capital) as given and examines how these differences affect trade, capital flows, and labour flows, and through them prices, wages, and returns to capital. In this model trade and factor mobility are substitute ways to reduce the economic effects of differing factor endowments and thus to reduce price and wage differences across countries. Restrictions of trade induce capital or labour flows that substitute for the restricted trade, and conversely restrictions on factor mobility induce trade (Mundell, 1957).

To analyse trade among countries with similar levels of economic development, economists use Ricardian models of trade. These models treat differences in technology as the fundamental determinant of trade and factor flows and examine how investments in technology create comparative advantage. Factor mobility magnifies differences in factor endowments because labour and capital move to economies where the technological advantage creates greater demand for them. Trade and factor mobility are complements in the sense that a technologically advantaged sector which uses, say highly skilled labour, will attract highly skilled immigrants to help it expand.

Because factor endowments and technology differ across countries and change over time, both sets of models are needed to make sense of globalization and labour.

When Factor Endowments Differ

If one identifies skilled labour, unskilled labour, capital, and natural resources as the relevant factors of production, trade patterns between advanced and developing countries fit the Heckscher–Ohlin model to a first approximation (Debaere, 2003, gives a favourable reading of the empirical validity of this model, while Trefler, 1995, is more critical). Countries with abundant skilled labour, such as the United States, export goods produced by skilled workers, and import goods made by low-skill labour, while countries with natural resources export those resources and import goods and services made with other inputs. But Heckscher–Ohlin models are silent on the huge volumes of trade among advanced countries with similar factor endowments and on the huge volumes of trade within industries (see Ruffin, 1999).

In addition, the pattern of factor flows is not consistent with the Heckscher–Ohlin model. Unskilled labour migrates from developing countries, where it is relatively abundant, to advanced countries, where it is relatively scarce, as the model predicts, but skilled labour also migrates to advanced countries, while it should move in the other direction. The brain drain, which is a sizable part of immigration, magnifies differences in factor proportions across countries, and thus creates a problem for analyses that view factor flows as responses to factor endowments.

The model also has predictions about the impact of globalization on factor prices that do not fit reality. It predicts that trade and factor flows will lower the relative returns to scarce factors and reduce the returns to abundant factors. This implies that globalization should increase wage differentials and inequality in advanced countries and reduce wage differentials and inequality in less advanced countries. Globalization is associated with rising inequality in advanced countries, which is consistent with the model, but it is also associated with rising inequality in many developing countries, which is inconsistent with the model. (For a review of studies showing rising wage differentials in developing countries associated with globalization, see Goldberg and Pavcnik, 2007.) One explanation for this is that the skilled workers in the developing countries are more comparable in their skills to the unskilled workers in the advanced countries, so that when the developing countries export products previously made by the unskilled workers in advanced countries, demand for skilled workers in developing countries is increased. But

other factors may be at work as well (see Zhu and Trefler, 2005), so there is no clear resolution to this surprising pattern.

Wage and Factor Price Equalization

Goods and factor flows motivated by national differences in factor endowments should reduce the cross-country dispersion of prices, the cost of capital and the wages of comparable workers. The factor price equalization theorem predicts that under specified conditions, trade alone will equalize factor prices. While some trade theorists dismiss factor price equalization as a theoretical curiosum, the logic of globalization dictates market pressures towards equality of wages as well as other prices across country lines.

In fact, the prices of many goods and services differ only moderately across countries. For instance, in 2004 the price of McDonald's Big Mac sandwich showed a narrow distribution across countries. The 80th percentile of Big Mac prices among 65 countries was 2.65 dollars while the 20th percentile of Big Mac prices was 1.40 dollars – a 1.9:1 spread (Freeman, 2006, p. 151). Similarly, estimates of international differences in the cost of capital show a ratio of costs at the top 25th percentile of countries to costs at the bottom 25th percentile of 1.43. This averages estimates from five different sources from Hail and Leuz (2004, Table 1). By contrast, the variation of wages in the same occupation is much greater. The 1998–2002 occupational wages around the world data file shows that wages for the country at the top 20 per cent point of the earnings distribution of countries for a *given narrowly defined* occupation are about 12 times the wages in the country at the bottom 20 per cent point of earnings distribution, if one uses exchange rates to compare currencies, and four to five to one if one compares currencies with purchasing power parity price indices – that is, price indices of different currencies based on a given basket of good that differ from exchange rates in part due to the different prices of non-tradables across countries (Freeman and Oostendorp, 2001). While part of the cross-country variation in wages for workers in the same occupation reflects differences in the education and skill of workers in the same occupation in advanced and developing countries, this cannot explain the wide variation in the earnings of, say, barbers in low-income countries and in high-income countries. The offshoring of computer programming and of call-centre work to India in the 2000s highlighted the fact that in some occupations workers in low-income countries have similar skills to those in more advanced countries. What differs are the wages paid across countries.

That the cross-country dispersion of wages is greater than the cross-country dispersion of prices or of the cost of capital suggests that globalization has had a smaller impact on the price of labour than on the prices of other factors. One possible reason for this is that, as noted earlier, international migration is a quantitatively smaller facet of globalization than trade or the international flow of capital. This explanation requires that the direct effect of trade on the prices of goods and services and the direct effect of capital flows on the cost of capital is greater than their indirect effect on wages.

The Labour Standards Debate

Globalization has made labour standards – workplace safety, freedom from discrimination, rights to unionize, hours and wage regulations – in developing countries a major issue for the international community. Human rights activists in advanced countries campaign to get multinational firms to implement better labour conditions in their plants and in those of their subcontractors in developing countries. The activists contend that consumers are willing to pay for the higher standards through higher prices and will avoid products made under bad conditions. There is indeed evidence that consumers will pay a bit more for 'fair trade' products and will shun products made under poor conditions (see Elliot and Freeman, 2003; Hiscox and Smyth, 2005). In response to activist pressures, many multinationals have developed and implemented codes of conduct for their operations in developing countries. Although activists fear that low standards in

developing countries will produce a global race to the bottom in standards, labour standards have risen in advanced countries during the period of rapidly increased trade with developing countries and in many developing countries as well. One indication of this is that advanced and developing countries have signed on to more International Labor Organization conventions during the period of globalization than ever before. Even the poorest countries have sought to reduce the use of child labour (see Elliot and Freeman, 2003).

Some advocates of free trade regard the activist pressure for improved labour standards in developing countries as disingenuous protectionism. 'The talk of "exploitation", failure to pay a "living wage" ... (is) little more than cynical manipulation of our moral instincts and an obfuscation of the reality to pursue our economic interest' (Bhagwati, 2000). 'The demand for linkage between trading rights and the observance of standards with respect to the environment and labour would seem to arise largely from protectionist motivation' (Srinivasan, 1994, p. 36). But the activists are not rival producers of imported products who aim to move production from developing countries to advanced countries. Rather, they are students, consumers, and members of non-governmental organizations who seek to organize retail markets so that consumers pay higher prices for items made under better conditions. Their motivation is intrinsic, not pecuniary interest.

What troubles free trade advocates is the danger that standards will impair the comparative advantage of developing countries. Motives aside, even policies intending to help workers in developing countries could harm them if those policies were so costly that they reduced the cost advantage of developing countries to expand in some low-wage labour-intensive sectors. However, the huge gap in labour costs between countries suggests that improved standards cannot threaten comparative advantage. In any case, part of the cost of standards falls on workers who prefer higher standards to lower standards; and part will be paid by consumers who want products made under good conditions. Some standards that raise costs to firms, moreover, benefit developing economies over the long run. Child labour laws, and school attendance laws, for instance, increase human capital formation; while occupational safety regulations reduce injuries and fatalities that may burden a country's medical or welfare system. As long as countries have flexible exchange rates, moreover, they can buy whatever labour standards they want without suffering economic disaster. If Brazil chooses to spend more on occupational health and safety than China, Brazilian firms will be at a competitive disadvantage *at a given exchange rate*. But the Brazilian currency will depreciate in relation to the Chinese currency, and all Brazilians will bear the cost of the health and safety standards through the higher cost of imports. Brazilian industries that spend a lot more to meet health and safety standards will contract as Brazil's comparative advantage shifts to industries that do not need to spend much more. Thus, globalization does not restrict national choices in labour standards or in other areas of social choice.

Globalization When Technology Differs

In a truly global labour market the same worker would earn roughly comparable real pay in different countries, as measured in the purchasing power parity price indices that give a more realistic comparison of living standards across countries than do comparisons of earnings based on exchange rates. The labour market at the outset of the 21st century was far from a single global market. Workers from a low-income country could make more than six times the earnings in their home country by immigrating to an advanced country (see Freeman, 2006). Why? Because the advanced country had higher capital–labour ratios, superior infrastructure, greater legal protections of property and persons, and more *advanced technology*, which raised productivity compared to productivity in the immigrant's native country. The differences in technology that affect earnings should be thought of in broad terms as including differences in organizational structure and business practices as well as differences in engineering or scientific technology. They include the economies of scale that give

'first mover' advantages to the firm or country that produces a good first. Countries with a comparative advantage in a sector – higher productivity in relation to other sectors compared to trading partners – will export output of that sector and import goods and services from sectors in which its trading partner has a comparative advantage. The sector with comparative advantage will expand, raising the wages of the factors that it uses the most and attracting those factors from the country and the rest of the world. If an advanced country has comparative advantage in, say, high tech that uses many computer scientists, persons with computer science degrees will immigrate to the country and strengthen its advantage in that sector. Trade and mobility magnify differences in factor endowments. Since countries will shift resources towards sectors in which they have comparative advantage, world output will rise, and so too will the wages of workers.

The 'North–South' model provides a platform for analysing trade between advanced countries (North) and developing countries (South) when investment in technology creates comparative advantage. In this model the North's advantage is in innovative high-tech products because it has many scientists, engineers, and other high-skilled workers, while the South's advantage is in producing standard products that use less skilled labour. The wages of ordinary workers in the North exceed those of workers in the South because the North earns a monopoly rent on technological innovation. The wage advantage is higher the greater the rate of technological innovation in relation to the rate of knowledge transfers to the South. The result is an industry or product life cycle that begins with an innovation in the North and ends with production in the South. Krugman (1979) gives a clear exposition of this model.

When technology creates comparative advantage, the productivity advances in one country can affect the economy of a trading partner positively or negatively depending on whether the advance occurs in goods or services that the trading partner exports or goods/services that it imports. If a trading partner improves productivity in an import, this will reduce the cost of production and the price of the import, which benefits the country that imports the good as well as (in most cases) the exporter. But if a trading partner improves technology in an export, this can harm the exporting country, just as an improved technology in a competitor can harm a firm. The increased supply of an exported good will drive down its price and thus the income of the country that originally dominated the production.

As a result, when countries make their comparative advantage by investing in skills or technology rather than having comparative advantage set by factor endowments there can be situations of 'conflicting national interests' in trade, as stressed by Gomory and Baumol (2000). If a foreign competitor gains comparative advantage in industries that have desirable attributes – that employ large numbers of highly educated and skilled workers or offer great opportunities for rapid technological advance – the lead economy will have to shift resources to less desirable sectors – and lose some of the advantages it had gained from trade. Applying these analyses to debates over offshoring and technological transfer in the United States, Paul Samuelson reminded trade economists that while the spread of technology around the world raises world output and productivity it need not be in the interest of the technological leader (Samuelson, 2004).

Globalization/Labour Debates: Human Resource Leapfrogging

The rapid growth of higher education in populous developing countries, notably China and India, challenges the assumption that advanced countries inevitably have comparative advantage in high-tech sectors. The share of scientific papers from Asia has risen substantially, due largely to increased scientific activity in China, which is moving to the forefront of science and technology. Digitalization of work has led to offshoring computer-related work, particularly to India. To take advantage of low-priced scientists and engineers in these countries, multinational firms have established research centres there. While the South has far fewer scientists and engineers per capita than the North, it can compete in high tech

because success at the technological frontier depends on the *absolute* number of scientists and engineers in an area, not simply on the number in relation to the total work force. A country like China or India can have proportionately fewer scientists, engineers, and entrepreneurs per capita than an advanced country but still have absolutely more of these workers available at lower wages than the advanced country. By producing numerous graduate scientists, engineers, and other university specialists and deploying them in the high-tech innovative sectors that the advanced countries had viewed as their birthright, the populous developing countries can move to the technological frontier through 'human resource leapfrogging'.

This does not mean that advanced countries lose when developing countries raise their technological prowess and economic competitiveness. The increased supply of highly educated workers around the world should expand the world's production possibility frontier rapidly, which will benefit all countries. In addition, the lower prices of high-tech goods and services produced in developing countries, such as PCs from China and call centre technical advice from India, benefit all consumers. But increased competition from low-wage countries in sectors where advanced countries have had comparative advantage can reduce or eliminate their advantages in those sectors. Comparative advantage in high-tech or most other sectors is not the birthright of any country. Globalization of technological progress will raise world output and income. It is likely to benefit workers in developing countries more than those in advanced countries, reducing global income inequality. It could lower the living standards or rate of growth of the living standards of advanced country workers for whom workers in low-income countries are good substitutes.

Does Globalization Rule the Roost? Will It Dominate Labour Outcomes in the Future?

In 1995 I posed the question 'Are your wages set in Beijing?' to direct attention towards the impact of globalization on wages in advanced countries (Freeman, 1995). My answer then, and now, is negative. The dispersion of wages for similar work around the world documented earlier shows that globalization does not rule labour markets. National labour markets, and the policies and institutions that unions, firms, and countries use to regulate those markets, affect wages and worker well-being independent of what happens in other countries. But the pressures of globalization on wage setting around the world will rise as the highly populous economies of China and India increase their share of the global economy. Globalization makes what happens in Beijing ... and Calcutta ... and Rio ... and Warsaw and so on, important drivers of labour market developments worldwide. Still, the persistence of variation in labour market outcomes across the states in the United States, where there are no restrictions on goods, factor flows, or knowledge flow, suggests that, even though global economic forces are likely to increase their impact on wages and other outcomes, they will not 'rule the roost'. There will remain space for variation in labour markets among countries just as there is for regional and local markets within countries.

See Also

- ▶ Globalization
- ▶ Heckscher–Ohlin trade theory
- ▶ Labour economics
- ▶ Purchasing power parity
- ▶ Ricardian trade theory
- ▶ Trade, technology diffusion and growth

Bibliography

Bhagwati, J. 2000. Why Nike is on the right track. *Financial Times*, 1 May.

Debaere, P. 2003. Relative factor abundance and trade. *Journal of Political Economy* 111: 589–610.

Elliot, K., and R. Freeman. 2003. *Can labor standards improve under globalization?* Washingon, DC: Institute for international economics.

Freeman, R.B. 1995. Are your wages set in Beijing? *Journal of Economic Perspectives* 9(3): 15–32.

Freeman, R. 2005a. *The great doubling: America in the new global economy.* Usery Lecture, Georgia State University, 8 April.
Freeman, R. 2005b. What really ails Europe (and America): The doubling of the global workforce. *The Globalist*, 3 June.
Freeman, R. 2006. People flows. *Journal of Economic Perspectives* 20(2): 145–170.
Freeman, R.B., and R.H. Oostendorp. 2001. The occupational wages around the world data file. *International Labour Review* 140: 379–401.
Goldberg, P.K., and N. Pavcnik. 2007. Distributional effects of globalization in developing countries. *Journal of Economic Literature* 45: 39–82.
Gomory, R., and W. Baumol. 2000. *Global trade and conflicting national interests*. Cambridge, MA: MIT Press.
Hail, L., and C. Leuz. 2004. International differences in the cost of equity capital: Do legal institutions and securities regulation matter? Working Paper 04–06, Wharton Financial Institutions Center, University of Pennsylvania. Online. Available at http://fic.wharton.upenn.edu/fic/papers/04/0406.pdf. Accessed 12 June 2007.
Hiscox, M., and N. Smyth. 2005. Is there consumer demand for improved labor standards? Evidence from field experiments in social labeling. Harvard University. Online. Available at http://www.courses.fas.harvard.edu:9095/Bgov3009/Calendar/SocialLabeling_2.pdf. Accessed 17 June 2007.
International Monetary Fund. 2007. World Economic Outlook database 2004. Online. Available at http://www.imf.org/external/pubs/ft/weo/2004/01/data/index.htm. Accessed 12 June 2007.
Krugman, P. 1979. A model of innovation, technology transfer, and the world distribution of income. *Journal of Political Economy* 87: 253–266.
Mundell, R. 1957. International trade and factor mobility. *American Economic Review* 47: 321–335.
Porter, A.L., J.D. Roessner, N. Newman, X.-Y. Jin, and D. Johnson. 2006. *High tech indicators: Technology-based competitiveness of 33 nations: 2005 final report*. Atlanta: Technology Policy and Assessment Center, Georgia Institute of Technology.
Ruffin, R.J. 1999. The nature and significance of intra-industry trade. *Economic and Financial Review.* Federal Reserve Bank of Dallas. Online. Available at http://www.dallasfed.org/research/efr/1999/efr9904a.pdf. Accessed 12 June 2007.
Samuelson, P. 2004. Where Ricardo and Mill Rebut and confirm arguments of mainstream economists supporting globalization. *Journal of Economic Perspectives* 18(3): 135–146.
Srinivasan, T.N. 1994. International labor standards once again! In *International Labor Standards and Global Economic Integration: Proceedings of a Symposium*, ed. G.K. Schoepfle and K.A. Swinnerton. Washington, DC: Bureau of International Labor Affairs, US Department of Labor.
Trefler, D. 1995. The case of the missing trade and other mysteries. *American Economic Review* 85: 1029–1046.
Zhu, S.C., and D. Trefler. 2005. Trade and inequality in developing countries: A general equilibrium analysis. *Journal of International Economics* 65: 21–48.

Globalization and the Welfare State

Heinrich W. Ursprung

Abstract

Most scholarly investigations do not support the often heard claim that globalization impairs the welfare state: observed cuts in welfare programmes appear to be mainly driven by domestic factors. The empirical evidence supporting the converse claim – that globalization gains are used to compensate the losers from global economic integration – is, however, also inconclusive. In order to disentangle the multifaceted and potentially inconsistent globalization effects on the plethora of welfare state activities, future research will have to adopt a more explicit micro-orientation, better econometric techniques, and an empirical research strategy that is more firmly based on political economic theories.

Keywords

Compensation effect of globalization; Deindustrialization; Diffusion processes; Efficiency effect of globalization; Globalization; Globalization and the welfare state; Labour market institutions; Labour market risk; Political economy; Risk sharing; Social insurance; Trade openness; Welfare state

JEL Classifications

F1

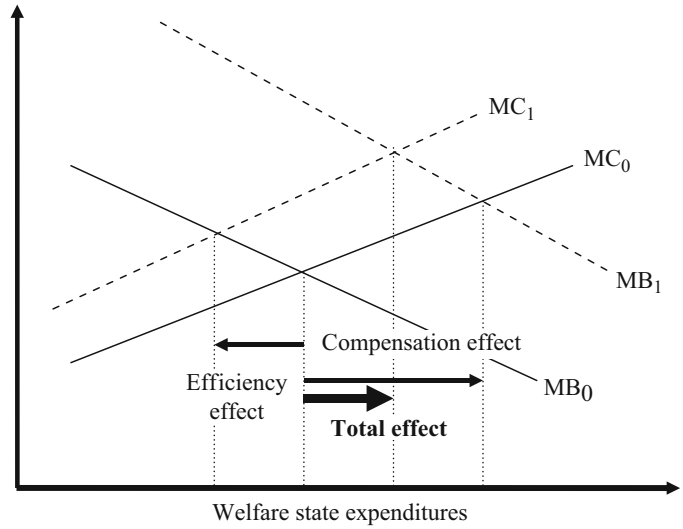
Globalization and the welfare state are commonly considered to be related; more precisely, many people believe that globalization exerts a negative

influence on the size and scope of the welfare state. This contention has been examined in an impressive number of scholarly investigations. Since globalization has far-reaching effects on income distribution, this issue has, however, attracted not only social scientists but also all kinds of political entrepreneurs: well-meaning public figures concerned with the globalization-induced social dynamic, political demagogues vying for political support, and even street rioters.

The worries of the well-meaning objectors to global economic integration originate in the conviction that globalization will bring about a loss of power of the nation states in general. They argue that liberalization of international transactions renders tax bases increasingly footloose, which induces a global tax race to the bottom and, as a consequence, jeopardizes the nation states' ability to finance welfare state activities. This downward pressure on the supply side of public welfare programmes, depending on the viewpoint of the observer, reduces the *efficiency* of benevolent governments and/or *disciplines* egoistic governments that transform discretionary power into political support. At any rate, the so-called efficiency or discipline effect of globalization tends to reduce the size and scope of government welfare programs.

The opponents of globalization ignore, however, the demand side of the political market, which derives from governments' political support maximization motives to redistribute the gains from globalization; that is, the losers from globalization, in particular workers who become exposed to higher labour market risks, will to some extent be *compensated* via an extension of social welfare programmes. Reviving Ruggie's (1982) notion of 'embedded liberalism', Rodrik (1998) interprets this kind of compensation as an exchange of social insurance in return for public support for openness. The so-called 'compensation' effect thus counteracts the 'efficiency' effect, implying that, from a theoretical point of view, the total effect of globalization on the welfare state remains ambiguous.

The interaction of the two effects is summarized in Fig. 1. The marginal benefit (in terms of political support) of welfare state expenditures decreases, whereas the marginal cost increases. Political support is maximized at the level where the MB and the MC curves intersect. A deepening or widening of economic integration now increases the marginal cost of supplying social welfare programmes and also increases the marginal benefit via increased demand, thereby shifting the two curves upwards (MC_0 to MC_1 and MB_0 to MB_1). Whether the resulting *efficiency effect* of globalization dominates the *compensation effect* or vice versa is a matter that can be resolved only by empirical research.



Globalization and the Welfare State, Fig. 1 Effects of globalization on social welfare expenditures

The first-generation literature on the globalization–welfare state nexus which appeared in the 1990s focused on this task of estimating globalization's net influence on government spending. Surveying this literature, Schulze and Ursprung (1999, pp. 345–7) arrive at the following conclusion:

> The general picture drawn by the few econometric studies available thus far does not lend any support to any alarmist view. At an aggregate level, many of these studies find no negative relationship between globalization and the nation states' ability to conduct independent fiscal policies. . . . no strong evidence points to a significant globalization-induced change of the level of public spending. But also accustomed expenditure patterns do not appear to have changed in the course of globalization. This may be due, however, to a lack of studies using strongly disaggregated public expenditure data.

In the meantime, many scholars have indeed taken up this implicit challenge and have used more disaggregated data; others have analysed specific groups of countries, have refined the empirical methods, or have investigated non-economic routes of influence. To foreshadow the main result of the second-generation literature, no unambiguous consensus has thus far emerged from these investigations. Instead, the new approaches have painted a multi-faceted picture that does not lend itself to a straightforward overall interpretation. Common to all empirical investigations is, however, the general research strategy which implies – usually in the framework of a panel data-set – the regression of some measure of welfare state activities on a set of explanatory variables which include globalization-related determinants, domestic economic and/or demographic determinants, and variables describing the domestic political–institutional setting.

As far as the dependent variable is concerned, one observes that measures of the size of the government sector (such as the ratio of total government expenditures to GDP) have been replaced by variables that better describe the size or scope of welfare policies; examples are expenditures on health, education, and social security (usually as a share of GDP) or net replacement rates of unemployment insurance. In addition, it has been argued that year-to-year policy adjustments and discontinuous policy shifts may be governed by different forces. Hicks and Zorn (2005), for example, identify welfare retrenchment events in OECD countries and find that they are not induced by direct globalization effects. They do, however, dampen subsequent globalization-induced increases of social spending. Factors that promote welfare policies may therefore (by doing so excessively and in ways that aggravate the policy's negative by-products) build up pressures for sudden policy reversals. Another novel way of looking at the welfare-state implications of globalization proposed by Hays et al. (2005) bridges the micro–macro divide in the traditional literature by investigating the extent to which trade policy stances of voters depend on the welfare policies conducted in their respective home countries. It transpires that protectionist sentiments of workers in import-competing industries can indeed be reduced with welfare policies that provide some kind of insurance against the labour market risk that is supposed to be associated with international exposure.

The crucial independent variables capturing the influence of globalization have also become more precisely tailored over time. Traditional measures of globalization, such as total international trade (imports plus exports) as a share of GDP, financial deregulation indices, and foreign direct investment (FDI) as a share of GDP, are not sufficiently focused on the dark side of globalization which gives rise to compensatory demands. Better suited for this purpose are, for example, imports as a share of GDP, imports from low-wage countries as a share of total imports, and outward FDI flows as a share of GDP. A further refinement of these measures is based on the argument that flow shares are biased with respect to country size because small countries are by nature more open than larger ones. Bretschger and Hettich (2002) have therefore decided to work with a measure of trade openness which controls for country size. A more substantial refinement consists in taking non-economic dimensions of globalization into account. Dreher (2006) derives indices of three dimensions of globalization – namely, economic, political and social globalization – and arrives at

the result that none of the three dimensions appears to have a significant impact on overall and social expenditures. Jahn (2006), in a similar attempt, introduces diffusion processes in order to capture globalization forces that transcend purely economic channels of influence. Diffusion mechanisms allow for learning and emulation processes which, in a setting of deepening globalization, become increasingly important since political-economic integration allows all political agents to better compare domestic policy efficiency with policy efficiency pursued in other countries. The diffusion variable employed in Jahn's study measures – for each country – the weighted average of the dependent variable of the other countries, where the weights represent closeness in terms of the respective bilateral trade intensity. This diffusion variable has a statistically significant influence on the social expenditure behaviour in the 16 OECD countries analysed, which may be interpreted to imply that globalization, via political yardstick competition, facilitates the adoption of best-practice policies.

Even though most of the second-generation studies do not support the efficiency hypothesis, there are notable exceptions. Garrett and Mitchell (2001), in particular, have claimed that 'year to year' increases in total trade are associated with less total government spending, less government consumption and lower security benefits as a share of GDP. This result has been construed to imply a preponderance of the efficiency effect over the compensation effect even though most of the financial globalization indicators employed in this particular study actually do not point into this direction and the share of low-wage country imports even appears to have a positive, compensatory, effect. More momentous than the daring interpretation of the study's results is, however, the critique that has been levelled against the econometric modelling. A whole series of recent studies have taken this controversial study as a starting point for scrutinizing the appropriate specification of panel data regression models of welfare state development (Kittel and Winner 2005; Plümper et al. 2005; Podesta 2006). What has unambiguously emerged from this deliberation is that panel data inferences react in a very sensitive manner to the chosen model specification and that the controversial Garrett–Mitchell estimates are driven by misspecifications. Re-estimating the relationship using the original variables but statistically better-behaved models reveals that changes in government spending are primarily driven by the state of the domestic economy.

This finding is in line with the influential study by Iversen and Cusack (2000), who do not find any relationship between globalization and the level of labour- market risk, presumably because trade not only increases this risk via specialization, but at the same time also diversifies it across a larger market, implying that the net effect of increased trade exposure remains ambiguous. According to Iversen and Cusack, most of the uncertainties in modern societies originate from dislocations caused by technologically induced structural transformations, and it is this transformation towards deindustrialization that has spurred electoral demands for welfare state compensation and risk sharing. Demand for welfare state extensions thus appear to be largely home-made.

Most studies investigating the nexus between globalization and the welfare state focus on OECD countries, that is, countries that were blessed with democratic government in the period under review. Only recently have scholars begun to concern themselves with developing countries, which requires that one squarely address the crucial role of the political regime and the labour market institutions in accommodating the demand for welfare policies. After all, defending welfare benefits under the pressures of globalization is likely to be much easier in countries honouring civil and political rights than in countries in which the losers of globalization are politically powerless and poorly organized. Most studies indeed confirm that democracies are more responsive to compensation demands than autocratic regimes. Avelino et al. (2005), to name a prominent study on Latin America, arrive at a somewhat more differentiated conclusion: they confirm that democracies have a strong positive influence on social spending, but also point out that it is questionable whether this influence is reinforced by globalization- induced

social insurance demands. Further analysis is needed to obtain a better understanding of how globalization and regime effects interact.

Even though the globalization–welfare state nexus has been the subject of intense empirical research since the early 1990s, the underlying relationship has remained rather elusive. To be sure, the second-generation studies corroborate the conclusions that have been drawn from the earlier studies, that is, the evidence certainly does not point towards an alarming globalization-induced 'race to the bottom'. The available signs of compensatory welfare policy measures are, however, rather weak and inconclusive. What does this mean in the final analysis? Some scholars believe that globalization simply does not matter. Swank (2002, pp. 119–20), for example, squarely declares that 'the conventionally hypothesized globalization dynamics are absent. Internationalization has no systematic impact on welfare policy change.' On the other hand, one could argue that globalization gives rise to significant efficiency and compensation effects which, however, neutralize each other. In order to discriminate between these two views the macro-perspective that characterizes the bulk of the relevant literature does not seem to be helpful. Future research, therefore, will have to undergo a reorientation.

The leading scholars in the field no longer believe that globalization and the welfare state represent low-dimensional phenomena which are linked by a simple causal relationship. The ambiguous empirical results are often attributed to the fact that the nexus between globalization and the welfare state is more complex than the mechanism illustrated in Fig. 1 suggests. This view implies that future research needs to focus even more on the interaction at the level of very specific globalization forces and on social policy responses which are just as narrowly specified. Nobody expects, however, that these micro-level interdependencies will add up to a sufficiently consistent macroeconomic response that would allow ostentatious claims of the sort that stimulated the early literature.

A second aspect that is likely to influence future research concerns econometric methods. To be sure, much progress has been made in perfecting empirical methods, but the struggle for better methods has certainly not yet come to an end. This is not meant to imply that the ambiguity of the results will disappear when proper empirical methods are applied across the board. On the other hand, it cannot be denied that methodological shortcomings may well be responsible for some of the observed discrepancies between studies that analyse closely related issues.

A final point concerns the fact that the empirical models employed almost invariably rest on ad hoc postulates. Even though political economy has become an integral part of mainstream economic thinking, theoretical arguments concerning the identity and the stakes of the involved political agents and a detailed investigation of how the strategic interaction of these interests shapes the ongoing political process have not found their way into the predominantly empirical literature on welfare state development. Investigations which are firmly based on political economy arguments would not only guide the empiricist in identifying worthwhile associations but also shift the focus of the discussion from addressing questions associated with political recriminations back to issues which are more securely rooted in the traditional scholarly discourse.

See Also

- ▶ Globalization
- ▶ Welfare State

Bibliography

Avelino, G., D. Brown, and W. Hunter. 2005. The effects of capital mobility, trade openness, and democracy on social spending in Latin America, 1980–1999. *American Journal of Political Science 49*: *625–641*.

Bretschger, L., and F. Hettich. 2002. Globalisation, capital mobility and tax competition: Theory and evidence for OECD countries. *European Journal of Political Economy* 18: 695–716.

Dreher, A. 2006. The influence of globalization on taxes and social policy – an empirical analysis for OECD countries. *European Journal of Political Economy* 22: 179–201.

Garrett, G., and D. Mitchell. 2001. Globalization, government spending and taxation in the OECD. *European Journal of Political Research* 39: 145–177.

Hays, J., D. Ehrlich, and C. Peinhardt. 2005. Government spending and public support for trade in the OECD: An empirical test of the embedded liberalism thesis. *International Organization* 59: 473–494.

Hicks, A., and C. Zorn. 2005. Economic globalization, the macro economy, and reversals of welfare: Expansion in affluent democracies, 1987–94. *International Organization* 59: 631–662.

Iversen, T., and T. Cusack. 2000. The causes of welfare state expansion: Deindustrialization or globalization? *World Politics* 52: 313–349.

Jahn, D. 2006. Globalization as 'Galton's Problem': The missing link in the analysis of diffusion patterns in welfare state development. *International Organization* 60: 401–431.

Kittel, B., and H. Winner. 2005. How reliable is pooled analysis in political economy? The globalization–welfare state nexus revisited. *European Journal of Political Research* 44: 269–293.

Plümper, T., P. Manow, and V. Troeger. 2005. Pooled data analysis in the comparative political economy of the welfare state: A note on methodology and theory. *European Journal of Political Research* 44: 327–354.

Podesta, F. 2006. Comparing time series cross-section model specifications: The case of welfare state development. *Quality & Quantity* 40: 539–559.

Rodrik, D. 1998. Why do more open economies have bigger governments? *Journal of Political Economy* 106: 997–1032.

Ruggie, J.G. 1982. International regimes, transactions, and change: Embedded liberalism in the postwar economic order. *International Organization* 36: 379–415.

Schulze, G., and H.W. Ursprung. 1999. Globalization of the economy and the nation state. *World Economy* 22: 295–352.

Swank, D. 2002. *Global capital, political institutions, and policy change in developed welfare states*. Cambridge: Cambridge University Press.

Godwin, William (1756–1836)

Peter Marshall

The first and greatest exponent of philosophical anarchism, Godwin was born in Wisbech, Cambridgeshire in 1756. He was brought up in the Dissenting tradition in Norfolk, attended Hoxton Academy, and became a candidate minister. He gradually lost his faith, and in his late twenties turned to political journalism for the Whig cause. Inspired by the French Revolution, he wrote *An Enquiry concerning Political Justice* (1793). It earned him immediate recognition: 'no work in our time', Hazlitt wrote, 'gave such a blow to the philosophic mind of the country'. His novel *Caleb Williams* (1794) was considered no less of a masterpiece. But as the reaction to the French Revolution grew, so Godwin's reputation waned. Despite a long series of grew, so Godwin's reputation waned. Despite a long series of novels, histories, plays, essays and children's books, he was unable to recapture the public imagination. He died in 1836, and was buried beside the feminist Mary Wollstonecraft, who had died in childbirth. Their daughter Mary eloped with Godwin's greatest disciple, Percy Bysshe Shelley.

Godwin's economics, like his politics, are an extension of his ethics. His starting point is a belief in the perfectibility of man: since man is a rational and voluntary being, education will suffice to make him enlightened, generous and free. In his ethics, he is a thoroughgoing and consistent utilitarian, defining good as pleasure and arguing that 'I should contribute everything in my power to the general good'. Indeed, his bold application of the principle of utility, coupled with the principle of impartiality, led him to condemn private affections, positive rights, promises, gratitude and patriotism. In his politics, he concluded that government and law are unnecessary evils and in their place proposed a decentralized and simplified society of autonomous communities.

Godwin considers the subject of property (or economics) as the keystone that completes the fabric of political justice. His treatment has a critical and a constructive phase. He sees a close link between property and power: the rich are always 'directly or indirectly the legislators of the state'. Moreover, accumulated property has disastrous effects on rich and poor alike: it creates a 'servile and truckling spirit', makes wealth the universal passion, and reduces society to the narrowest selfishness.

But since we have a common nature, it follows from the principle of impartial justice that the good things of the world are a 'common stock', upon which one man is as entitled as another to draw what he wants. Property therefore should be considered a trust to be employed in the best possible way in order to promote liberty, knowledge and virtue. Just as every man has a duty to help his neighbour, so his neighbour has a claim to assistance.

Developing the labour theory of value, Godwin further argues that money is only a means of exchange and that there is no wealth except the labour of man. The producer should therefore retain what is necessary for his subsistence from the produce of his labour and then distribute the surplus to the most needy. Godwin also distinguishes between four classes of things: the means of subsistence, the means of intellectual and moral improvement, inexpensive pleasures, and luxuries. It is the last which is the chief obstacle to the just distribution of the previous three.

In place of the capitalist economic system, Godwin looks to small-scale production for the local market in a decentralized society. Production would be organized voluntarily with the producers controlling distribution. There would be a voluntary sharing of material goods, without barter or exchange. Anticipating the liberating effects of new technology, Godwin suggests that if all able-bodied people worked, production time could be reduced drastically, thereby giving people the leisure to develop their intellectual and moral potential.

When Malthus asserted that such a scheme would result in over-population, Godwin replied with his doctrine of moral restraint or prudence as a check. In his *Of Population* (1820), he went on to question the validity of Malthus's ratios, and argued that people would tend to reproduce less as their living standards improved.

Godwin's economic theory is clearly both profound and original. He was the first to write systematically about the competing claims of capital, need and production. Marx and Engels recognized his importance in developing the theory of exploitation. He not only strongly influenced the early socialist thinkers Robert Owen, William Thompson and Thomas Hodgskin, but the Owenites and Chartists took note of what he had to say. While Malthus has been most remembered, it is arguable that Godwin will be proved right in the long run. His scheme of voluntary communism remains moreover a thoughtful and persuasive ideal.

Selected Works

1793. An enquiry concerning political justice and its influence on general virtue and happiness, 2 vols. London: Robinson.
1794. Things as they are: or, they adventures of Caleb Williams, 3 vols. London: B. Crosby.
1797. The enquirer: Reflections on education, manners and literature. London: Robinson.
1801. Thoughts occasioned by the perusal of Dr. Parr's Spital Sermon. London: Robinson.
1820. Of population. An enquiry concerning the power of increase in the numbers of mankind. London: Longman, Hurst, Rees, Orme & Brown.
1831. Thoughts on man, his nature. Productions and discoveries. London: Wilson.

References

Brailsford, H.N. 1913. *Shelley, Godwin and their circle*. Oxford: Oxford University Press.
Clark, J.P. 1977. *The philosophical anarchism of William Godwin*. Princeton: Princeton University Press.
Hazlitt, W. 1820. William Godwin. In *The spirit of the age*. Oxford: Oxford University Press, 1954.
Locke, D. 1980. *A fantasy of reason: The life and thought of William Godwin*. London/Boston/Henley: Routledge & Kegan Paul.
Marshall, P.H. 1984. *William Godwin*. New Haven/London: Yale University Press.
Monro, D.H. 1953. *Godwin's moral philosophy*. Oxford: Oxford University Press.
Paul, C.K. 1876. *William Godwin: His friends and contemporaries*, vol. 2. London: H.S. King.
Priestley, F.E.L. 1946. *Introduction to enquiry concerning political justice*, vol. III. Toronto: University of Toronto Press.

Gold Standard

Lawrence H. Officer

Abstract

The world has had two experiences with gold standards: the classical gold standard and the interwar gold standard. The 'rules of the game', government policies to preserve the gold standard, were rarely followed. Rather, government responsible policy and credible commitment to the standard, private stabilizing arbitrage and speculation, and stable political and economic environment made the classical gold standard a success. The absence of these elements and the presence of the Great Depression combined to make the interwar gold standard a failure.

Keywords

Arbitrage; Bank of France; Bank rate; Bank Restriction Period; Banking crises; Bimetallism; Bullion; Bullion standard; Central banking; Commitment; Convertibility; Current account deficits; Deflation; Exchange controls; Federal Reserve System; Fixed exchange rates; Floating exchange rates; Gold standard; Gold-exchange standard; Great Depression; Monetary base; Money multiplier; Money supply; Orthodox metallism; Reichsbank; Reserve ratio; Silver standard; Specie-flow mechanism; Speculation; Sterilization; Sticky prices; Sticky wages

JEL Classifications

N2

The classical gold standard (which ended in 1914) and the interwar gold standard are examined within the same framework, but their experiences are vastly different.

Types of Gold Standard

All gold standards involve (*a*) a fixed gold content of the domestic monetary unit, and (*b*) the monetary authority both buying and selling gold at the mint price (the inverse of the gold content of the monetary unit), whereupon the mint price governs in the marketplace. A 'coin' standard has gold coin circulating as money. Privately owned bullion (gold in form other than domestic coin) is convertible into gold coin, at (approximately) the mint price, at the government mint or central bank. Private parties may melt domestic coin into bullion – the effect is as if coin were sold to the monetary authority for bullion. The authority could sell gold bars directly for coin, saving the cost of coining.

Under a pure coin standard, gold is the only money. Under a mixed standard, there are also notes issued by the government, central bank, or commercial banks, and possibly demand deposits. Government or central-bank notes (and central-bank deposit liabilities) are directly convertible into gold coin at the fixed price on demand. Commercial-bank notes and demand deposits are convertible into gold or into gold-convertible government or central-bank currency. Gold coin is always exchangeable for paper currency or deposits at the mint price. Two-way transactions again fix the currency price of gold at the mint price.

The coin standard, naturally 'domestic', becomes 'international' with freedom of international gold flows and of foreign-exchange transactions. Then the fixed mint prices of countries on the gold standard imply a fixed exchange rate (mint parity) between their currencies.

A 'bullion' standard is purely international. Gold coin is not money; the monetary authority buys or sells gold bars for its notes. Similarly, a 'gold-exchange' standard involves the monetary authority buying and selling not gold but rather gold-convertible foreign exchange (the currency of a country on a gold coin or bullion standard).

For countries on an international gold standard, costs of importing and exporting gold give rise to

'gold points', and therefore a 'gold-point spread', around the mint parity. If the exchange rate, number of units of domestic per unit of foreign currency, is greater (less) than the gold export (import) point, arbitrageurs sell (purchase) foreign currency at the exchange rate and also obtain (relinquish) foreign currency by exporting (importing) gold. The domestic-currency cost of the transaction per unit of foreign currency is the gold export (import) point; so the 'gold-point arbitrageurs' receive a profit proportional to the exchange-rate/gold-point divergence. However, the arbitrageurs' supply of (demand for) foreign currency returns the exchange rate to below (above) the gold export (import) point. Therefore perfect arbitrage would keep the exchange rate within the gold-point spread. What induces gold-point arbitrage is the profit motive and *the credibility of the monetary-authorities' commitment* to (*a*) the fixed gold price and (*b*) freedom of gold and foreign-exchange transactions.

A country can be effectively on a gold standard even though its legal standard is bimetallism. This happens if the gold – silver mint-price ratio is greater than the world price ratio. In contrast, even though a country is legally on a gold standard, its government and banks could 'suspend specie payments', that is, refuse to convert their notes into gold; so that the country is in fact on a 'paper standard'.

Countries on the Classical Gold Standard

Britain, France, Germany and the United States were the 'core countries' of the gold standard. Britain was the 'centre country', indispensable to the spread and functioning of the standard. Legally bimetallic from the mid-13th century, Britain switched to an effective gold standard early in the 18th century. The gold standard was formally adopted in 1816, ironically during a paper-standard regime (Bank Restriction Period). The United States was legally bimetallic from 1786 and on an effective gold standard from 1834, with a legal gold standard established in 1873–4 – also during a paper standard (the greenback period). In 1879 the United States went back to gold, and by that year not only the core countries but also some British dominions and non-core western European countries were on the gold standard. As time went on, a large number of other countries throughout the globe adopted gold; but they (along with the dominions) were in 'the periphery' – acted on rather than actors – and generally (except for the dominions) not as committed to the gold standard.

Almost all countries were on a mixed coin standard. Some periphery countries were on a gold-exchange standard, usually because they were colonies or territories of a country on a coin standard.

In 1913, the only countries not on gold were traditional silver – standard countries (Abyssinia, China, French Indochina, Hong Kong, Honduras, Morocco, Persia, Salvador), some Latin American paper-standard countries (Chile, Colombia, Guatemala, Haiti, Paraguay), and Portugal and Italy (which had left gold but 'shadowed' the gold standard, pursuing policies as if they were gold-standard countries, keeping the exchange rate relatively stable).

Elements of Instability in Classical Gold Standard

Three factors made for instability of the classical gold standard. First, the use of foreign exchange as official reserves increased as the gold standard progressed. While by 1913 only Germany among the core countries held any measurable amount of foreign exchange, the percentage for the rest of the world was double that for Germany. If there were a rush to cash in foreign exchange for gold, reduction of the gold of reserve-currency countries would place the gold standard in jeopardy.

Second, Britain was in a particularly sensitive situation. In 1913, almost half of world foreign-exchange reserves was in sterling, but the Bank of England had only three per cent of gold reserves. The Bank of England's 'reserve ratio'

(ratio of 'official reserves' to 'liabilities to foreign monetary authorities held in London financial institutions') was only 31 per cent, far lower than those of the monetary authorities of the other core countries. An official run on sterling could force Britain off the gold standard. Private foreigners also held considerable liquid assets in London, and could themselves initiate a run on sterling.

Third, the United States was a source of instability to the gold standard. Its Treasury held a high percentage of world gold reserves (in 1913, more than that of the three other core countries combined). With no central bank and a decentralized banking system, financial crises were more frequent and more severe than in the other core countries. Far from the United States assisting Britain, gold often flowed from the Bank of England to the United States, to satisfy increases in US demand for money. In many years the United States was a net importer rather than exporter of capital to the rest of the world – the opposite of the other core countries. The political power of silver interests and recurrent financial panics led to imperfect credibility in the US commitment to the gold standard. Indeed, runs on banks and on the Treasury gold reserve placed the US gold standard near collapse in the 1890s. The credibility of the Treasury's commitment to the gold standard was shaken; twice the US gold standard was saved only by cooperative action of the Treasury and a bankers' syndicate, which stemmed gold exports.

Automatic Force for Stability: Price Specie-Flow Mechanism

The money supply is the product of the money multiplier and the monetary base. The monetary authority alters the monetary base by changing its gold holdings and domestic assets (loans, discounts, and securities). However, the level of its domestic assets is dependent on its gold reserves, because the authority generates demand liabilities (notes and deposits) by increasing its assets, and convertibility of these liabilities must be supported by a gold reserve. Therefore the gold standard provides a constraint on the level (or growth) of the money supply.

Further, balance-of-payments surpluses (deficits) are settled by gold imports (exports) at the gold import (export) point. The change in the money supply is the product of the money multiplier and the gold flow, providing the monetary authority does not change its domestic assets. For a country on a gold-exchange standard, holdings of foreign exchange (a reserve currency) take the place of gold.

A country experiencing a balance-of-payments deficit loses gold and its money supply decreases *automatically.* Money income contracts and the price level falls, thereby increasing exports and decreasing imports. Similarly, a surplus country gains gold, exports decrease, and imports increase. In each case, balance-of-payments equilibrium is restored via the current account, the 'price specie-flow mechanism'. To the extent that wages and prices are inflexible, movements of real income in the same direction as money income occur; the deficit country suffers unemployment, while the payments imbalance is corrected.

The capital account also acts to restore balance, via interest-rate increases in the deficit country inducing a net inflow of capital. The interest-rate increases also reduce real investment and thence real income and imports. The opposite occurs in the surplus country.

Rules of the Game

Central banks were supposed to reinforce (rather than 'sterilize') the effect of gold flows on the monetary base, thereby enhancing the price specie-flow mechanism. A gold outflow decreases the international assets of the central bank and the money supply. The central-bank's 'proper' response is: (1) decrease lending and sell securities, thereby decreasing domestic assets and the monetary base; (2) raise its 'discount rate', which induces commercial banks to adopt a higher reserves–deposit ratio, thereby reducing the money multiplier. On both counts, the money supply is further decreased. Should the central

bank increase its domestic assets when it loses gold, it engages in sterilization of the gold flow, violating the 'rules of the game'. The argument also holds for gold inflow, with sterilization involving the central bank decreasing its domestic assets when it gains gold.

Monetarist theory suggests the 'rules' were inconsequential. Under fixed exchange rates, gold flows adjust money supply to money demand; the money supply is not determined by policy. Also, prices, interest rates, and incomes are determined worldwide. Even core countries can influence these variables domestically only to the extent that they help determine them in the global marketplace. Therefore the price-specie flow and like mechanisms cannot occur. Historical data support this conclusion: gold flows were too small to be suggestive of these processes; and, at least among the core countries, prices, incomes, and interest rates moved closely in correspondence, contradicting the specie-flow mechanism and rules of the game.

Rather than rule (1), central-bank domestic and international assets moving in the same direction, the opposite behaviour – sterilization – was dominant, both in core and non-core European countries. The Bank of England followed the rule more than any other central bank, but even so violated it more often than not!

The Bank of England did, in effect, manage its discount rate ('Bank Rate') in accordance with rule (2). The Bank's primary objective was to maintain convertibility of its notes into gold, and its principal tool was Bank Rate. When the Bank's 'liquidity ratio' (ratio of gold reserves to outstanding note liabilities) decreased, it usually increased Bank Rate. The increase in Bank Rate carried with it market short-term interest rates, inducing a short-term capital inflow and thereby moving the exchange rate away from the gold-export point. The converse also held, with a rise in the liquidity ratio generating a Bank Rate decrease. The Bank was constantly monitoring its liquidity ratio, and in response altered Bank Rate almost 200 times over 1880–1913.

While the Reichsbank also generally moved its discount rate inversely to its liquidity ratio, other central banks often violated rule (2). Discount-rate changes were of inappropriate direction, or of insufficient magnitude or frequency. The Bank of France kept its discount rate stable, choosing to have large gold reserves, with payments imbalances accommodated by fluctuations in its gold rather than financed by short-term capital flows. The United States, lacking a central bank, had no discount rate to use as a policy instrument.

Reason for Stability: Credible Commitment to Convertibility

From the late 1870s onward, there was absolute private-sector credibility in the commitment to the fixed domestic-currency price of gold on the part of Britain, France, Germany, and other important European countries. For the United States, this absolute credibility applied from about 1900. That commitment had a contingency aspect: convertibility could be suspended in the event of dire emergency; but, after normal conditions were restored, convertibility and honouring of gold contracts would be re-established at the pre-existing mint price – even if substantial deflation was required to do so. The Bank Restriction and greenback periods were applications of the contingency. From 1879, the 'contingency clause' was exercised by none of these countries.

The absolute credibility in countries' commitment to convertibility at the existing mint price implied that there was zero 'convertibility risk' (Treasury or central-bank notes non-redeemable in gold at the established mint price) and zero 'exchange risk' (alteration of mint parity, institution of exchange control, or prohibition of gold export).

Why was the commitment to credibility so credible?

1. Contracts were expressed in gold; abandonment of convertibility meant violation of contracts – anathema to monetary authorities.
2. Shocks to economies were infrequent and generally mild.
3. The London capital market was the largest, most open, most diversified in the world, and its gold market was also dominant. A high

proportion of world trade was financed in sterling, London was the most important reserve-currency centre, and payments imbalances were often settled by transferring sterling assets rather than gold. Sterling was an international currency – a boon to other countries, because sterling involved positive interest return, and its transfer costs were much less than those of gold. Advantages to Britain were the charges for services as an international banker, differential interest return on its financial intermediation, and the practice of countries on a sterling (gold-exchange) standard of financing payments surpluses with Britain by piling up short-term sterling assets rather than demanding Bank gold.

4. 'Orthodox metallism' – authorities' commitment to an anti-inflation, balanced-budget, stable-money policy – reigned. This ideology implied low government spending, low taxes, and limited monetization of government debt. Therefore, it was not expected that a country's price level would get out of line with that of other countries.
5. Politically, gold had won over paper and silver, and stable-money interests (bankers, manufacturers, merchants, professionals, creditors, urban groups) over inflationary interests (farmers, landowners, miners, debtors, rural groups).
6. There was a competitive environment and freedom from government regulation. Prices and wages were flexible. The core countries had virtually no capital controls, Britain had adopted free trade, and the other core countries had only moderate tariffs. Balance-of-payments financing and adjustment were without serious impediments.
7. With internal balance an unimportant goal of policy, preservation of convertibility of paper currency into gold was the primary policy objective. Sterilization of gold flows, though frequent, was more 'meeting the needs of trade' (passive monetary policy) than fighting unemployment (active monetary policy).
8. The gradual establishment of mint prices over time ensured that mint parities were in line with relative price levels; so countries joined the gold standard with exchange rates in equilibrium.
9. Current-account and capital-account imbalances tended to be offsetting for the core countries. A trade deficit induced a gold loss and a higher interest rate, attracting a capital inflow and reducing capital outflow. The capital-exporting core countries could stop a gold loss simply by reducing lending abroad.

Implications of Credible Commitment

Private parties reduced the need for balance-of-payments adjustment, via both gold-point arbitrage and stabilizing speculation. When the exchange rate was outside the spread, gold-point arbitrage quickly returned it to the spread. Within the spread, as the exchange value of a currency weakened, the exchange rate approaching the gold-export point, speculators had an ever greater incentive to purchase domestic with foreign currency (a capital inflow). They believed that the exchange rate would move in the opposite direction, enabling reversal of their transaction at a profit. Similarly, a strengthened currency involved a capital outflow. The further the exchange rate moved toward a gold point, the greater the potential profit opportunity in betting on a reversal of direction; for there was a decreased distance to that gold point and an increased distance from the other point. This 'stabilizing speculation' increased the exchange value of depreciating currencies, and thus gold loss could be prevented. Absence of controls meant such private capital flows were highly responsive to exchange-rate changes.

Government Policies That Enhanced Stability

Specific government policies enhanced gold-standard stability. First, by the turn of the 20th century, South Africa – the main world gold producer – was selling all its gold output in London, either to private parties or to the Bank of England. Thus the Bank had the means to

replenish its gold reserves. Second, the orthodox-metallism ideology and the leadership of the Bank of England kept countries' monetary policies disciplined and in harmony. Third, the US Treasury and the central banks of the other core countries manipulated gold points, to stem gold outflow. The cost of exporting gold was artificially increased (for example, by increasing selling prices for bars and foreign coin) and/or the cost of importing gold artificially decreased (for example, by providing interest-free loans to gold importers).

Fourth, central-bank cooperation was forthcoming during financial crises. The precarious liquidity position of the Bank of England meant that it was more often the recipient than the provider of financial assistance. In crises, the Bank would obtain loans from other central banks, and the Bank of France would sometimes purchase sterling to support that currency. When needed, assistance went from the Bank of England to other central banks. Also, private bankers unhesitatingly made loans to central banks in difficulty.

Thus, 'virtuous' interactions were responsible for the stability of the gold standard. The credible commitment to convertibility of paper money at the established mint price, and therefore to fixed mint parities, were both a cause and an effect of the stable environment in which the gold standard operated, the stabilizing behaviour of arbitrageurs and speculators, and the responsible policies of the authorities – and these three elements interacted positively among themselves.

Experience of Periphery

An important reason for periphery countries to join and maintain the gold standard was the fostering of access to core-countries' capital markets. Adherence to the gold standard connoted that the peripheral country would follow responsible macroeconomic policies and repay debt. This 'seal of approval', by reducing the risk premium, involved a lower interest rate on the country's bonds sold abroad, and very likely a higher volume of borrowing, thereby enhancing economic development.

However, periphery countries bore the brunt of the burden of adjustment of payments imbalances with the core (and other western European) countries. First, when the gold-exchange-standard periphery countries ran a surplus (deficit), they increased (decreased) their liquid balances in the United Kingdom (or other reserve-currency country) rather than withdraw gold from (lose gold to) the reserve-currency country. The monetary base of the periphery country increased (decreased), but that of the reserve-currency country remained unchanged. Therefore, changes in domestic variables – prices, incomes, interest rates, portfolios – that occurred to correct the imbalance were primarily in the periphery.

Second, when Bank Rate increased, London drew funds from France and Germany, which attracted funds from other European countries, which drew capital from the periphery. Also, it was easy for a core country to correct a deficit by reducing lending to, or bringing capital home from, the periphery. While the periphery was better off with access to capital, its welfare gain was reduced by the instability of capital import. Third, periphery-countries' exports were largely primary products, sensitive to world market conditions. This feature made adjustment in the periphery take the form more of real than financial correction.

The experience of adherence to the gold standard differed among periphery groups. The important British dominions and colonies successfully maintained the gold standard. They paid the price of serving as an economic cushion to the Bank of England's financial situation; but, compared with the rest of the periphery, gained a stable long-term capital inflow. In southern Europe and Latin America, adherence to the gold standard was fragile. The commitment to convertibility lacked credibility, and resort to a paper standard occurred. Many of the reasons for credible commitment that applied to the core countries were absent. There were powerful inflationary interests, strong balance-of-payments shocks, and rudimentary banking sectors. The cost of adhering to the gold standard was apparent: loss of the ability to depreciate the currency to counter reductions in exports. Yet the gain, in

terms of a steady capital inflow from the core countries, was not as stable or reliable as for the British dominions and colonies.

Breakdown of Classical Gold Standard

The classical gold standard was at its height at the end of 1913, ironically just before it came to an end. The proximate cause of the breakdown of the classical gold standard was the First World War. However, it was the gold-exchange standard and the Bank of England's precarious liquidity position that were the underlying cause. With the outbreak of war, a run on sterling led Britain to impose extreme exchange control – a postponement of both domestic and international payments – making the international gold standard inoperative. Convertibility was not suspended legally; but moral suasion, legalistic action, and regulation had the same effect. The Bank of England commandeered gold imports and applied moral suasion to bankers and bullion brokers to restrict gold exports.

The other gold-standard countries undertook similar policies – the United States not until 1917, when it adopted extra-legal restrictions on convertibility and restricted gold exports. Commercial banks converted their notes and deposits only into currency. Currency convertibility made mint parities ineffective; floating exchange rates resulted.

Return to the Gold Standard

After the First World War, a general return to gold occurred; but the interwar gold standard differed institutionally from the classical gold standard. First, the new gold standard was led by the United States, not Britain. The US embargo on gold exports was removed in 1919, and currency convertibility at the pre-war mint price was restored in 1922. The gold value of the dollar rather than pound sterling was the typical reference point around which other currencies were aligned and stabilized. The core now had two central countries, the United Kingdom (which restored gold in 1925) and the United States.

Second, for many countries there was a time lag between stabilizing the currency in the foreign-exchange market (fixing the exchange rate or mint parity) and resuming currency convertibility. The interwar gold standard was at its height at the end of 1928, after all core countries were fully on the standard and before the Great Depression began. The only countries that never joined the interwar gold standard were the USSR, silver-standard countries (China, Hong Kong, Indochina, Persia, Eritrea), and some minor Asian and African countries.

Third, the 'contingency clause' of convertibility conversion, that required restoration of convertibility at the mint price that existed prior to the emergency (the First World War), was *broken* by various countries, and even core countries. While some countries (including the United States and United Kingdom) stabilized their currencies at the pre-war mint price, others (including France) established a gold content of their currency that was a fraction of the pre-war level: the currency was devalued in terms of gold, the mint price was higher than pre-war. Still others (including Germany) stabilized new currencies adopted after hyperinflation.

Fourth, the gold coin standard, dominant in the classical period, was far less prevalent in the interwar period. All four core countries had been on coin in the classical gold standard; but only the United States was on coin interwar. The goldbullion standard, non-existent pre-war, was adopted by the United Kingdom and France. Germany and most non-core countries were on a gold-exchange standard.

Instability of Interwar Gold Standard

The interwar gold standard was replete with forces making for *instability.*

1. The process of establishing fixed exchange rates was piecemeal and haphazard, resulting in disequilibrium exchange rates. Among core countries, the United Kingdom restored convertibility at the pre-war mint price without sufficient deflation, and had an overvalued

currency of about ten per cent. France and Germany had undervalued currencies.

2. Wages and prices were less flexible than in the pre-war period.
3. Higher trade barriers than pre-war also restrained adjustment.
4. The gold-exchange standard economized on total world gold via the gold of the United Kingdom and United States in their reserves role for countries on the gold-exchange standard and also for countries on a coin or bullion standard that elected to hold part of their reserves in London or New York. However, the gold-exchange standard was unstable, with a conflict between (*a*) the expansion of sterling and dollar liabilities to foreign central banks, to expand world liquidity, and (*b*) the resulting deterioration in the reserve ratio of US and UK authorities.

   This instability was particularly severe, for several reasons. First, France was now a large official holder of sterling, and France was resentful of the United Kingdom. Second, many more countries were on the gold-exchange standard than pre-war. Third, the gold-exchange standard, associated with colonies in the classical period, was considered a system inferior to a coin standard.
5. In the classical period, London was the one dominant financial centre; in the interwar period it was joined by New York and, in the late 1920s, Paris. Private and official holdings of foreign currency could shift among the two or three centres, as interest-rate differentials and confidence levels changed.
6. There was maldistribution of gold. In 1928, official reserve-currency liabilities were much more concentrated than in 1913, British pounds accounting for 77 per cent of world foreign-exchange reserves and French francs less than two per cent (versus 47 and 30 per cent in 1913). Yet the United Kingdom held only seven per cent of world official gold and France 13 per cent. France also possessed 39 per cent of world official foreign exchange. The United States held 37 per cent of world official gold.
7. Britain's financial position was even more precarious than in the classical period. In 1928, the gold and dollar reserves of the Bank of England covered only one-third of London's liquid liabilities to official foreigners, a ratio hardly greater than in 1913. UK liquid liabilities were concentrated on stronger countries (France, United States), whereas UK liquid assets were predominantly in weaker countries (Germany). There was ongoing tension with France, which resented the sterling-dominated gold-exchange standard and desired to cash in its sterling holding for gold, to aid its objective of achieving first-class financial status for Paris.
8. Internal balance was an important goal of policy, which hindered balance-of-payments adjustment, and monetary policy was influenced by domestic politics rather than geared to preservation of currency convertibility.
9. Credibility in authorities' commitment to the gold standard was not absolute. Convertibility risk and exchange risk could be high, and currency speculation could be destabilizing rather than stabilizing. When a country's currency approached or reached its gold-export point, speculators might anticipate that currency convertibility would not be maintained and that the currency would be devalued.
10. The 'rules of the game' were violated even more often than in the classical gold standard. Sterilization of gold inflows by the Bank of England can be viewed as an attempt to correct the overvalued pound by means of deflation. However, the US and French sterilization of their persistent gold inflows reflected exclusive concern for the domestic economy and placed the burden of adjustment (deflation) on other countries.
11. The Bank of England did not provide a leadership role in any important way, and central-bank cooperation was insufficient to establish credibility in the commitment to currency convertibility. The Federal Reserve had three targets for its discount-rate policy: strengthen the pound, combat speculation in the New York stock market, and achieve internal balance – and the first target was of lowest priority. Although, for the sake of external

balance, the Bank of England kept Bank Rate higher than internal considerations would dictate, it was understandably reluctant to abdicate Bank Rate policy entirely to the balance of payments, with little help from the Federal Reserve. To keep the pound strong, substantial international cooperation was required, but was not forthcoming.

Breakdown of Interwar Gold Standard

The Great Depression triggered the unravelling of the gold standard. The depression began in the periphery. Low export prices and debt-service requirements created insurmountable balance-of-payments difficulties for gold-standard commodity producers. However, US monetary policy was an important catalyst. In 1927 the Federal Reserve favoured easy money, which supported foreign currencies but also fed the New York stock-market boom. Reversing policy to tame the boom, higher interest rates attracted monies to New York, weakening sterling in particular. The crash of October 1929, while helping sterling, was followed by the US depression.

This spread worldwide, with declines in US trade and lending. In 1929 and 1930 a number of periphery countries –both dominions and Latin American countries – either formally suspended currency convertibility or restricted it so that currencies violated the gold-export point.

It was destabilizing speculation, emanating from lack of confidence in authorities' commitment to currency convertibility, which ended the interwar gold standard. In May 1931 there was a run on Austria's largest commercial bank, and the bank failed. The run spread to other eastern European countries and to Germany, where an important bank also collapsed. The countries' central banks lost substantial reserves; international financial assistance was too late; and in July 1931 Germany adopted exchange control, followed by Austria in October. These countries were definitively off the gold standard.

The Austrian and German experiences, as well as British budgetary and political difficulties, were among the factors that destroyed confidence in sterling, which occurred in mid-July 1931. Runs on sterling ensued, and the Bank of England lost much of its reserves. Loans from abroad were insufficient, and in any event taken as a sign of weakness. The gold standard was abandoned in September, and the pound quickly and sharply depreciated on the foreign-exchange market, as overvaluation of the pound would imply.

Following the UK abandonment of the gold standard, many countries followed, some to maintain their competitiveness via currency devaluation, others in response to destabilizing capital flows. The United States held on until 1933, when both domestic and foreign demands for gold, manifested in runs on US commercial banks, became intolerable. 'Gold bloc' countries (France, Belgium, Netherlands, Switzerland, Italy, Poland), with their currencies now overvalued and susceptible to destabilizing speculation, succumbed to the inevitable by the end of 1936.

The Great Depression was worsened by the gold standard: gold-standard countries hesitated to inflate their economies, for fear of suffering loss of gold and foreign-exchange reserves, and being forced to abandon convertibility or the gold parity. The gold standard involved 'golden fetters', which inhibited monetary and fiscal policy to fight the Depression. As countries left the gold standard, removal of monetary and fiscal policy from their 'gold fetters' enabled their use in expanding real output, providing the political will existed.

In contrast to the interwar gold standard, the classical gold standard functioned well because of a confluence of 'virtuous' interactions, involving government policies, credible commitment to the standard, private arbitrage and speculation, and fostering economic and political environment. We will not see its like again.

See Also

- ▶ Bank of England
- ▶ Banking Crises
- ▶ Bimetallism
- ▶ Bretton Woods System
- ▶ Commodity Money

- ▶ Silver Standard
- ▶ Specie-Flow Mechanism

Bibliography

Bayoumi, T., B. Eichengreen, and M.P. Taylor, eds. 1996. *Modern perspectives on the gold standard*. Cambridge: Cambridge University Press.

Bordo, M.D., and F.E. Kydland. 1995. The gold standard as a rule: An essay in exploration. *Explorations in Economic History* 32: 423–464.

Bordo, M.D., and H. Rockoff. 1996. The gold standard as a 'good housekeeping seal of approval'. *Journal of Economic History* 56: 389–428.

Bordo, M.D., and A.J. Schwartz, eds. 1984. *A retrospective on the classical gold standard, 1821–1931*. Chicago: University of Chicago Press.

De Macedo, J.B., B. Eichengreen, and J. Reis, eds. 1996. *Currency convertibility: The gold standard and beyond*. London: Routledge.

Eichengreen, B. 1992. *Golden fetters: The gold standard and the Great Depression, 1919–1939*. New York: Oxford University Press.

Eichengreen, B., and M. Flandreau. 1997. *The gold standard in theory and history*. 2nd ed. London: Routledge.

Gallarotti, G.M. 1995. *The anatomy of an international monetary regime: The classical gold standard, 1880–1914*. New York: Oxford University Press.

Officer, L.H. 1996. *Between the dollar-sterling gold points*. Cambridge: Cambridge University Press.

Officer, L.H. 2001. Gold standard. In *EH. Net encyclopedia*, ed. R. Whaples. Online. Available at http://eh.net/encyclopedia/article/officer.gold.standard. Accessed 20 Oct 2006.

Golden Age

Amit Bhaduri

The notion of equilibrium or steady state in economics differs from that in mechanics because, unlike particles and molecules, economic agents are guided in their action by expectations of the future. Moreover, expectations are often being falsified by actual events forcing the economic actors to revise continuously and adapt their expectations in the light of experience. However, in an economic equilibrium expectations are *never* falsified; what happens must be compatible with what people expected to happen in every period of time and all expectations are being continuously fulfilled. Since all expectations cannot always be fulfilled in actual history, this in itself should be an adequate warning about the utter implausibility of the notion of equilibrium in economics. Economic equilibrium is something that we can never observe in reality; at best, it has to be recognized as a 'thought experiment' designed to facilitate analysis (Robinson 1979).

An economically valid 'thought experiment' on the properties of steady state (or equilibrium) growth in a capitalist economy must then entail an analysis of how the various economic agents guided by their expectations, behave along such a path. This, in essence, would distinguish an economically meaningful description of steady state growth from its mere mechanical analogy in terms of a set of balance equations. Joan Robinson coined the term 'golden age' to demonstrate that, once expectations are explicitly dealt with, steady state growth represents an almost 'mythical state of affairs not likely to obtain in any actual economy' (Robinson 1956, p. 99).

The mechanical balance condition for the 'golden age' is the equality over time between the natural, the warranted and the actual rate of growth of national income ($g_n = g_w = g_a$) in the sense of Harrod (1948). This, in itself, represents a highly stringent set of conditions that are satisfied only accidentally. For instance, unless labour productivity rises at a uniform rate in all the sectors (in a multisectoral context) and the real (and individual product) wage rate(s) also rises at that same rate, technical progress would tend to upset the system of relative prices as well as the class distribution of income between wage and profit ruling along any steady growth path. Thus, technical progress, unless it is 'neutral' in the above sense, would not be compatible with golden age growth. But this could only be *accidentally* true, if one accepts that labour productivity growth through technical progress has a largely autonomous character.

Along the golden age, a constant set of relative prices rule and there is a uniform rate of

profit (r) in all sectors, given by the Cambridge formula $r = g_{n/}s_p$, where s_p is the propensity to save out of profit and, g_n, is the natural rate of growth (Robinson 1956; Pasinetti 1962). However, this *realized* profit rate must also be the *expected* profit rate ($\hat{\gamma}$) of the capitalists. Given that investment decisions are *autonomously* taken by the capitalists in the light of their profit expectations, at that expected rate of profit ($\hat{\gamma}$), the capitalists must autonomously decide to carry out just sufficient investment required at the golden age of growth (hence, $g_w = g_a$; Robinson 1962; Marglin 1984). Similarly, one can also presume that, with powerful trade unions, their *expected* real wage rate must always correspond to its actual path along the golden age. Thus, unless real wage is expected to rise at the same rate as labour productivity under neutral technical progress, collective wage bargain may divert the economy from the mythically tranquil conditions of social harmony represented along the golden age.

But even all this may be inadequate for maintaining the golden age. By its very nature, the equality between the natural, warranted and actual growth ($g_n = g_w = g_a$) ensures only a *flow equilibrium*, for example on this growth path the *additional* demand created by the multiplier through rising investment equals *additional* supply generated through net investment in each period. However, it does *not* ensure that any under- or over-utilization of the historically inherited capacities of the initial period would be alleviated. An *arbitrary* initial condition, say in terms of under-utilization of initial capacities or large initial unemployment, may therefore jeopardize the possibility of the golden age. This leads to its most paradoxical aspect: the initial conditions have to be suited exactly to the requirements of equilibrium growth rather than being arbitrary to sustain a golden age, that is, the economy must already be in equilibrium to continue along it!

See Also

► Robinson, Joan Violet (1903–1983)

Bibliography

Harrod, R. 1948. *Towards a dynamic economics*. London: Macmillan.

Marglin, S. 1984. Growth, distribution and inflation: A centennial synthesis. *Cambridge Journal of Economics* 8(2): 115–144.

Pasinetti, L.L. 1962. Rate of profit and income distribution in relation to the rate of economic growth. *Review of Economic Studies* 29: 267–279.

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

Robinson, J. 1962. A simple model of accumulation. In *Essays in the theory of economic growth*, ed. J. Robinson. London: Macmillan.

Robinson, J. 1979. History vs. equilibrium. In *Collected economic papers*, vol. 5, ed. J. Robinson. Oxford: Basil Blackwell.

Golden Rule

Edmund S. Phelps

Abstract

The so-called golden rule (of capital accumulation) is a proposition about the consequences for national welfare possibilities of alternative paths of national wealth, and hence of national saving, in a closed economy. It states that the steady-growth state that gives the maximum path of consumption is the one along which national consumption equals the national wage bill and thus national saving equals 'profits'. The basic significance of the golden rule is as a warning against national policies of over-saving or counterproductive austerity.

Keywords

Capital accumulation; Capital-output ratio; Golden rule; Maximin; National consumption function; Natural rate of growth; Phelps, E. S.; Reciprocity; Robinson, J. V.; Saving-output ratio; Schumpeter, J. A.; Social optimum; Social rate of return; Solow, R. M.; Steady-growth state; Swan, T. W.; Technological progress; Utilitarianism

JEL Classifications
D9

The so-called golden rule, or golden rule of capital accumulation, is a proposition about the consequences for national consumption – more broadly, for national welfare possibilities – of alternative paths of national wealth, and hence of national saving, in a closed economy. It developed out of the dynamic models of capital accumulation and output growth, generally in a setting of steady technical progress and demographic increase, begun in 1956 by R.M. Solow and T.W. Swan after some early explorations by Harrod, Domar and Robinson. Solow and Swan had shown that, provided diminishing returns to capital set in strongly enough (whether or not smoothly), there exists a state of steady growth corresponding to each possible value of the saving–output ratio, and, interestingly enough, this steady-state growth rate is independent of the value of the saving ratio – and often called the natural rate of growth. Hence, an upward shift of saving at each level of output, through increased private thrift or else higher taxes or lower spending by the government, cannot have a permanent, or non-vanishing, effect on the growth rate of output, only a transient effect.

If this theorem was the first law of the new 'growth economics', the golden rule of accumulation was the second law of growth economics. It states that the steady- growth state that gives the maximum path of consumption – the path layered on top of all the other steady-growth consumption tracks – is the one along which national consumption equals the national wage bill and thus national saving equals 'profits' (gross of interest in the present use of the term). Equivalently, the consumption- maximizing steady-growth path is the steady state along which the competitive rate of interest, which is the social rate of return to investment and to saving, is equal to the natural rate of growth. (To see the equivalence, divide profits and saving by capital.) Hence, a country (with any given history) that now plans for ever to equate saving to profits could not hope to achieve a sustainable increase in the consumption path by some date in the future through a shift of policy towards increased saving; very possibly, even a temporary increase of consumption would not result. The reason is that despite a boost to future output brought by greater accumulation, the increase in future investment would eat up the increase in future output – and then some.

The arrival *circa* 1960 of the golden rule result was a classic case of multiple discoverers. And discovery seems the apt word, since the golden rule theorem was just a simple insight about a set or sets of equations in existence for several years that was waiting to be noticed, not a creative vision of the world springing from an independent empirical sense; accordingly, many or most of the discoverers were fledgling, pre-flight theorists still working on the ground of existing models. The earliest publishers of the result were Phelps (1961), Robinson (1962) and Swan (1963). However, it quickly became apparent that there were also discoveries on the Continent by von Weizsäcker and Allais, and even within the tiny space of the Cowles Foundation at Yale there were additional independent discoveries by Beckmann and Srinivasan. Robinson coined the proposition 'the neoclassical theorem', but eventually Phelps's coinage 'the golden rule' became the standard. This was not a case of bad money driving out good, as will be explained.

The term 'golden rule' was something of a play on words. Mrs. Robinson had dubbed states of steady growth as 'golden ages', so a proposition (if not exactly a maxim) about choosing among golden ages was natural to call a golden rule. In addition, there was also an allusion in the term to the biblical golden rule, do unto others as you would have them do unto you. The sense of that maxim, presumably, is that if one asserts a right to a certain policy, or treatment, from others, then in one's own treatment of others one must accord them the right to the same policy; so the choice of the rights to assert is subject to a reciprocity or cost constraint, which is a useful thing, for otherwise one would demand the most extreme sacrifices of others. Of course, this precept – the national saving policy, or national consumption function, that a future generation would have preceding ones follow, in view of its self-interests, it

must likewise adopt on behalf of succeeding generations – does not by itself determine the just policy of national saving. Yet the golden rule perspective serves to alert us that there will be a limit to the austerity that future generations would ask of the present generation if they are obliged to practice the same austerity that they choose to preach. To make this effective, it should be noted, the saving policies from which society is to choose must be linear-homogeneous, and thus expressed in terms of saving as a ratio to output or profits or some related variable. (Otherwise, a future generation could piously call for lower consumption only at the present, comparatively low level of national income – and thus travel as a free rider.)

The meaning and indeed the significance of the golden rule becomes quite transparent in the special case of an economy in which the technology and the (working-age) population are constant, so that the natural rate of growth of the economy is zero. In this case the golden rule state is the Schumpeterian zero-interest stationary state; and since the net rate of return to investment is zero, gross profits are simply depreciation allowances and equal to gross investment, which is entirely replacement investment in a stationary state. Here it is abundantly clear why an alternative stationary state with a constant negative rate of interest would actually yield a lower path of consumption: the extra replacement investment would more than eat up the extra gross output, leaving an actual diminution of the net national product and consumption (to which NNP is equal). It is also perfectly clear that a society is not required as a matter of efficiency to aim for the Schumpeterian state; if the initial rate of return to investment is positive, it would take *lower* consumption in the present than would otherwise be possible on a sustainable basis (simply by consuming income) in order to move to the Schumpeterian state so that in the future a higher consumption level could be sustained than would otherwise be possible. Neither is such a move required as a matter of justice. From the utilitarian side there are economists who cheerfully discount the utilities of future people, and from the 'maximin' perspective it is obvious that present people would not optimally sacrifice to make better-off those who were not worse-off than they to begin with. The basic significance of the golden rule, then, is as a warning against national policies of over-saving, or counterproductive austerity. The golden rule theorem is simply a generalization to a growing economy of these observations.

Further results on the inefficiency entailed by exceeding, so to speak, the golden rule in certain respects were later obtained. It was shown with the help of T.C. Koopmans that keeping the capital–output ratio indefinitely in excess (by a nonvanishing amount) of the golden rule level would be dominated in terms of consumption, and thus utility, by another path, feasible from the same initial conditions, along which the capital–output ratio is always 'epsilon' smaller (Phelps 1965, 1966). A much more general analysis came later from D. Cass in 1972 in which the borderline between efficient and dynamically inefficient paths is systematically examined.

'But the golden rule path could be the social optimum, couldn't it? Certainly it is very beautiful, and not obviously unjust!' There was a tendency among some to regard it as the optimum at least provisionally, for working purposes. However, any budding claims that may have existed for the 'optimality' of the golden rule path met with an objection by I.F. Pearce (1962). Start there at T_1, Pearce said, and end there at T_2. Then, if there is steady population growth, so the golden rule interest rate is positive (being equal to the population growth rate), it will increase the integral of total utility to save more now and less later, causing the capital stock to arch over its golden-rule track, since more saving will increase output. There could be no denying this, although some utilitarians prefer to sum the per capita utility of people over time (or the utility of per capita consumption), which suggests there is a maximin impulse in their otherwise utilitarian hearts; from this angle it would not be preferable to deviate along Pearce's arching detour. Then, using the per capita utility version of utilitarianism, P.A. Samuelson (1967) took up the cudgels with a revision of the Pearce argument: if there is steady *technical progress,* so the golden rule interest rate exceeds the population growth rate, it will

increase the integral of per capita utility to cause the capital stock to arch above its golden-rule track since more saving will increase per capita output – as long as the interest rate remains above Samuelson's 'biological' level, which is the population growth rate. Again, it is *nolo contendere* from the golden rule side. Yet maximin advocates might object that if the per capita utility received by succeeding generations is rising, and unavoidably so, so that the oldest generation extant is always the worst-off, the detour from the golden rule path proposed by Samuelson would presumably entail some belt-tightening by the oldest generation along with the others in order to produce the consumption splurge for the benefit of some younger or future generations – hence a reduction in *minimum* per capita utility across generations. That cannot be a maximin improvement and is indeed a maximin worsening. Thus goes the maximin rejoinder to the turnpike 'refutation' of the golden rule.

See Also

- ▶ Neoclassical Growth Theory
- ▶ Ramsey Model
- ▶ Robinson, Joan Violet (1903–1983)

Bibliography

Cass, D. 1972. On capital overaccumulation in the aggregative neoclassical model of economic growth: A complete characterization. *Journal of Economic Theory* 4: 200–223.

Pearce, I.F. 1962. The end of the golden age in Solovia. *American Economic Review* 52: 1088–1097.

Phelps, E.S. 1961. The golden rule of accumulation: A fable for growthmen. *American Economic Review* 51: 638–643.

Phelps, E.S. 1965. Second essay on the golden rule of accumulation. *American Economic Review* 55: 793–814.

Phelps, E.S. 1966. *Golden rules of economic growth*. New York/Amsterdam: Norton.

Robinson, J. 1962. A neo-classical theorem. *Review of Economic Studies* 29: 219–226.

Samuelson, P.A. 1967. A turnpike refutation of the golden rule in a welfare-maximizing many-year plan. In *Essays on the theory of optimal economic growth*, ed. K. Shell. Cambridge, MA: MIT Press.

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

Swan, T.W. 1956. Economic growth and capital accumulation. *Economic Record* 32: 334–361.

Swan, T.W. 1963. Of golden ages and production functions. In *Economic development with special reference to East Asia*, ed. K.E. Berrill. London: Macmillan.

Goldfeld, Stephen (1940–1995)

Richard E. Quandt

Keywords

Computational economics; Disequilibrium models; Econometrics; Goldfeld, S.; GRADX algorithm; Homoscedasticity; Money demand function; Nonlinear models in econometrics; Sargan distribution; Suits agricultural model; Voice

JEL Classifications

B31

Goldfeld was born in Bronx, New York, and received his undergraduate training in mathematics at Harvard University. He obtained a Ph.D. in economics from MIT in 1963. Except for visiting appointments at Center for Operations Research and Econometrics (CORE) at the Université Catholique de Louvain, the University of California, Berkeley, and the Technion in Haifa, and a year as a member of the US Council of Economic Advisors, he spent his professional life at Princeton University, where he served as Provost from 1993 until his death.

Goldfeld largely divided his professional interests between monetary economics and econometric methodology. In the former category are his early book (1966a) and two definitive papers (1973a) and (1976) that are models of careful empirical econometrics. They address the question of whether the money demand function is basically stable in the short run. In estimating the

money demand function for the post-war period, Goldfeld exhaustively analyses specifications differing from one another in terms of the extent of aggregation, the type of lag structure and the types of variables included in the demand function. In (1973a), he finds no substantial short-run instability in the demand function. He also concludes that disaggregation by sector probably helps explain money demand. But the (1976) paper attacks the poor predictive performance of the money demand equation in the period following 1973: the equation consistently over-predicts money demand. An even more extensive and thorough analysis aimed at fixing this problem makes him suggest that a more fundamental rethinking of money demand is indicated. He also studied savings and loan associations and their behaviour with respect to rate setting and allocational efficiency (1970), forecasting and policy evaluation in large-scale econometric models (1971) and numerous related topics.

His methodological interests began with the construction of a new test for homoscedasticity (1965), subsequently called the Goldfeld–Quandt test, followed in short order by an abiding interest in nonlinear estimation (1968) and computational econometrics (1966b). This led to the invention of the GRADX algorithm, which solves the maximization problem for non-concave functions by embedding them in a concave analogue. He then turned to the case of switching regressions: problems in which the parameters of a regression equation switch from one set of values to another, but at an unknown point in time (1973b) (hence making the Chow test inappropriate). The superficial similarity of these models to disequilibrium models, in which the observed quantity represents the short side of the market, led him to an abiding interest in models of the latter type, sometimes within the framework of financial institutions (1975, 1980). A reconsideration of the Suits agricultural model may contain the first proof of the unboundednes of a likelihood function for a disequilibrium model when sample separation is not known a priori (1975). Because of the relatively intractable nature of the econometrics when error terms are normally distributed in these models, he explored the properties of a new distribution called the Sargan distribution (1981).

In the last few years of his life, Goldfeld was much interested in the microtheoretic problems of socialist enterprises operating in the presence of a soft budget constraint, a concept derived from János Kornai (1979, 1980, 1982). In the models, firms can undertake 'whining' when they face losses and bail-outs permit them to continue operating. These models confirm Kornai's conjecture that the expectation of bail-outs will lead to larger input demand than in the standard competitive case (1988, 1990). Embedding the possibility of 'whining' in a multi-firm case in which inputs are rationed and whining is intended to induce a more generous input allotment leads to the determination of a Nash equilibrium in whining (1990).

Goldfeld had enormous range and creativity and was a precise, careful and methodical worker as well as a much-loved teacher of both undergraduates and graduate students. His premature death was a great loss to the profession.

See Also

- ▶ Command Economy
- ▶ Money
- ▶ Numerical Optimization Methods in Economics

Selected Works

1965. (With R. Quandt.) Some tests for homoscedasticity. *Journal of the American Statistical Association* 60: 539–547.

1966a. *Commercial bank behavior and economic activity: A structural study of monetary policy in the postwar United States*. Contributions to Economic Analysis Series No. 43. Amsterdam: North-Holland.

1966b. (With R. Quandt and H. Trotter.) Maximization by quadratic hill-climbing. *Econometrica* 34: 541–551.

1968. (With R. Quandt.) Nonlinear simultaneous equations: Estimation and prediction. *International Economic Review* 91: 113–136.

1970. (With D Jaffee.) The determinants of deposit-rate setting by savings and loan associations. *Journal of Finance* 25: 615–632.
1971. Forecasting and policy evaluation using large scale econometric models: Comment. In *Frontiers of quantitative economics*, ed. M. Intriligator. Amsterdam: North-Holland.
1972. (With R. Quandt.) *Nonlinear methods in econometrics*. Contributions to Economic Analysis Series No. 77. Amsterdam: North-Holland.
1973a. The demand for money revisited. *Brookings Papers on Economic Activity* 1973: 3: 577–638.
1973b. (With R. Quandt.) The estimation of structural shifts by switching regressions. *Annals of Economic and Social Measurement* 2: 475–485.
1974a. (With D. Aigner.) Estimation and prediction from aggregate data when aggregates are measured more accurately than their components. *Econometrica* 42: 113–134.
1975. (With R. Quandt.) Estimation in a disequilibrium model and the value of information. *Journal of Econometrics* 3: 325–348.
1976. The case of the missing money. *Brookings Papers on Economic Activity* 1976: 3: 683–730.
1980. (With D. Jaffe and R. Quandt.) A Model of FHLBB advances: rationing or market clearing. *Review of Economics and Statistics* 62: 339–347.
1981. (With R. Quandt.) Econometrics modelling with nonnormal disturbances. *Journal of Econometrics* 17: 141–155.
1988. (With R. Quandt.) Budget constraints, bailouts and the firm under central planning. *Journal of Comparative Economics* 12: 502–520.
1990. (With R. Quandt.) Output targets, the soft budget constraint and the firm under central planning. *Journal of Economic Behavior and Optimization* 14: 205–222.
1994. (With R. Quandt.) The competition for rationed resources. *Journal of Economic Behavior and Organization* 25: 53–71.

Bibliography

Kornai, J. 1979. Resource-constrained versus demand-constrained systems. *Econometrica* 47: 801–820.
Kornai, J. 1980. *Economics of shortage*. Amsterdam: North-Holland.
Kornai, J. 1982. *Growth, shortage and efficiency*. Amsterdam: North-Holland.

Goldsmith, Raymond William (1904–1988)

E. Denison

Keywords
Financial intermediation ratios; Financial interrelations ratios; Goldsmith, R. W.; National accounting; National income; Perpetual inventory method; Reproducibility

JEL Classifications
B31

Born in Brussels, Goldsmith studied in Berlin (Ph. D., 1927). He emigrated to the United States and served on the US Securities and Exchange Commission (1934–41), the War Production Board (1942–7) and as a member of the Senior Staff, National Bureau of Economic Research (1953–78). He was a Professor at New York University (1956–61) and at Yale University (1962–73; Emeritus, 1973 onwards). He pioneered the measurement of national and sector saving, investment, wealth, and balance sheets.

In the 1930s Goldsmith initiated Securities and Exchange Commission estimates of saving, obtained by a new method: subtracting additions to liabilities from additions to assets. During 1947–68 the Commerce Department national income accounts reported personal saving derived from the Securities and Exchange Commission series as an alternative to estimates obtained as income less consumption and as investment less corporate and government saving. The Federal

Reserve Board then absorbed the Securities and Exchange Commission work into its flow-of-funds accounts. These had been started by Copeland in research partly paralleling Goldsmith's (Copeland 1952, p. xvi).

A still more important breakthrough, in 1950, was Goldsmith's perpetual inventory method of estimating the stock and consumption of durable capital. The United States and other countries use it to obtain official national income and capital stock series. Goldsmith's monumental *Study of Saving* followed, providing balance sheets, saving, and wealth, for the nation and sectors. It featured broad scope, great detail, and series running back to the 1890s (to 1805 for reproducible wealth; in Goldsmith 1951b). Goldsmith subsequently updated these series and introduced similar ones for other countries. Lifelong adherence to the principle of reproducibility, which calls for sufficient description of all estimates to enable the reader to reproduce them, increases the usefulness of Goldsmith's data.

Goldsmith wrote extensively about financial institutions and capital markets. He introduced the financial interrelations and financial intermediation ratios (Goldsmith 1966, 1969). In his later years Goldsmith studied pre-industrial economies (Goldsmith 1987).

Selected Works

1939. (Assisted by W. Salant.) The volume and components of saving in the United States 1933–1937. *Studies in income and wealth*, vol. 3. New York: NBER.

1951a. A perpetual inventory of national wealth. *Studies in Income and Wealth*, vol. 14. New York: NBER.

1951b. The growth of reproducible wealth of the United States of America from 1805 to 1950. In *Income and wealth of the United States: Trends and structure*, ed. S. Kuznets. Cambridge: Bowes and Bowes.

1955–6. *A study of saving in the United States*, 3 vols. Princeton: Princeton University Press.

1958. *Financial intermediaries in the American economy since 1900*. Princeton: Princeton University Press.

1962. *The national wealth of the United States in the postwar period*. Princeton: Princeton University Press.

1963. (With R. Lipsey and M. Mendelson.) *Studies in the national balance sheet of the United States*, 2 vols. Princeton: Princeton University Press.

1966. *The determinants of financial structure*. Paris: OECD.

1969. *Financial structure and development*. New Haven: Yale University Press.

1982. *The national balance sheet of the United States, 1953–1980*. Chicago: University of Chicago Press.

1984. An estimate of the size and structure of the national product of the Early Roman Empire. *Review of Income and Wealth* 30: 263–288.

1985. *Comparative national balance sheets: A study of twenty countries, 1688–1978*. Chicago: University of Chicago Press.

1987. *Premodern financial systems: A historical comparative study*. Cambridge: Cambridge University Press.

Bibliography

Copeland, M. 1952. *A study of money-flows in the United States*. New York: NBER.

Goldsmith, Selma (née Selma Evelyn Fine) (1912–1962)

E. Denison

A government economic statistician specializing in distributions of consumer units by size of income, Selma Goldsmith studied at Cornell (BA, 1932) and Radcliffe (PhD, 1936), and was the wife of the economist Raymond W. Goldsmith.

After participating in income estimation at the National Resources Committee and the Agriculture Department, Goldsmith initiated the Office of Business Economics size distributions.

Distributions were prepared annually as an adjunct to the office's national income accounts, and older distributions were adjusted for comparability. Before this work was suspended, Goldsmith and her successors published before-tax distributions of personal income in current and constant dollars for 20 of the years from 1929 through 1963 and after-tax distributions for 1950–62 (Fitzwilliam 1964).

Estimators elsewhere had supplemented surveys of consumer units (which contain large reporting errors) with information for upper income groups from tax returns, and adjusted resulting distributions to conform to independent totals for income and units. Goldsmith's innovations included starting with earnings from tax returns for the whole nonfarm distribution, using survey information only for farm units, to fill gaps, and to combine tax payers into family groups; adopting OBE's personal income concept; and constructing time series. Ingenious techniques, often devised by H. Kaitz, and meticulous use of all information characterized the distributions. However, data grouped by class intervals, rather than microfiles based on statistical or exact matching of tax and survey data, usually had to be used in combining sources or adjusting to control totals.

Goldsmith also compiled capital gains statistics (Seltzer 1951). At her death she was Chief of the Income and Statistics Branch of the Bureau of the Census.

See Also

► Social Accounting

Selected Works

1939. Fine, S. (With E. Baird.) The use of income tax data in the National Resources Committee estimate of the distribution of income by size. *Studies in income and wealth,* vol. 3. New York: National Bureau of Economic Research, 149–203.

1951a. Appraisal of basic data available for constructing income size distributions. *Studies in income and wealth,* vol. 13. New York: National Bureau of Economic Research, 266–373.

1951b. Seltzer, L., assisted by Goldsmith, S., and Kendrick, M. *The nature and tax treatment of capital gains and losses.* New York: National Bureau of Economic Research.

1954. (With G. Jaszi, H. Kaitz and M. Liebenberg.) Size distribution of income since the mid-thirties. *Review of Economics and Statistics* 36, 1–32.

1955a. Income distribution in the United States, 1950–53. *Survey of Current Business* 35(3), 14–27.

1955b. (With G. Jaszi, assisted by H. Kaitz and M. Liebenberg.) *Income distribution in the United States, a supplement to the survey of current business.*

1958. The relation of census income distribution statistics to other income data. (With comment by J. Pechman.) *Studies in Income and Wealth,* vol. 23. 65–113.

1959. Income distribution by size – 1955–58. *Survey of Current Business* 39(4), 9–16.

1960. Size distribution of personal income, 1956–59. *Survey of Current Business* 40(4), 8–15.

Bibliography

Fitzwilliam, J. 1964. Size distribution of income in 1963. *Survey of Current Business* 44(4): 3–11.

Gonner, Edward Carter Kersey (1862–1922)

Murray Milgate and Alastair Alastair

The character, causes and consequences of that long historical process which witnessed the transformation of British agriculture from a system of open field cultivation based on entitlements and obligations fixed by the custom of the manor, to a

system of large-scale enclosed farming characterized by modern relations between wage-labour and capital with the ownership of the land vested in private hands, constitutes perhaps the most difficult, controversial and fascinating set of historical questions facing the economist. From the time of the agrarian disturbances of the sixteenth century, popular opinion has associated the process with physical deprivation among the agricultural population, the depopulation of the countryside and, later, with the throwing of whole populations into urban centres where they were forced to subsist under conditions of severe economic hardship. While scholarly opinion on the subject has not proved to be so single-minded, there are a few contributions to it that seem likely to have a long life; Gonner's *Common Land and Inclosure* (1912a) is one of them.

Published in the same year as Tawney's *Agrarian Problem of the Sixteenth Century* and less than a year after the Hammond's *Village Labourer*, both of which dealt with the same subject, Gonner sets out to avoid 'general approval or condemnation' of varying views (p. v) by addressing the question using a descriptive and statistical account of actual movements and their effects. Of course, the basic scholarly positions had been set down in the preceding century in the writing of Marx, on the one hand, and that of authors like Cunningham on the other. The former supported a conflictual interpretation: Marx saw the movement as one of 'bloody legislation' which tore peasants from the soil and created a propertyless proletariat. Cunningham, however, rejected this view and saw the process as a more voluntaristic one essential to economic progress – enclosure was in a sense in everyone's interest. He also invoked the demographic transition to account for the origins of the industrial workforce. In an important way, the work of the Hammonds and Tawney represents an attempt to re-state the conflictual position using more detailed historical analysis than Marx (or, for that matter, Cunningham) had invoked. Though they qualified the orthodox Marxian vision in crucial ways, not least in terms of their rejection of the simplistic explanation which entailed the idea of an immanent contradiction between the social relations and the forces of production which dominated Marxist doctrine, they called into question the more optimistic position taken by Cunningham and others. Gonner's account seems to fall somewhere in between. He distinguishes between necessary and unnecessary effects and is on the whole convinced that the dislocation that did occur was an unnecessary consequence. He finds no evidence of any radical decline in employment in agriculture, but like Tawney he emphasizes the importance of the penetration of market relations into medieval agriculture over a very long period of time.

Gonner's early works include a tract against socialism (1895) and a study of the social philosophy of Rodbertus (1899). With his *Interest and Saving* of 1906, however, he turned his attention to a topic more susceptible to empirical analysis, and found there a mode of analysis more suited to his talents.

In addition to this, Gonner edited two volumes of Ricardo's works which are still quite well known: the *Principles* in 1891, and the *Economic Essays* published posthumously in 1923. To judge from his introductions to these volumes, Gonner seems to have had very little deep regard for the work of Ricardo. His verdict on the *Principles* is harsh, holding that 'not only is it remarkable for infelicity of language, with all its fatal consequences of exaggeration and obscurity, but the grammar itself is halting and the accuracy often apparent, fallaciously apparent rather than real' (p. xxiv). In the much later volume of Ricardo's *Economic Essays* while apparently more generous in finding Ricardo to be 'in reality far less abstract and far more inductive than many of his critics have allowed' (p. xviii), Gonner adds immediately a rough jibe at Ricardo's 'overstrained and incorrect' assumptions. He also attacks Ricardo's analysis of the relationship between the rate of profit and accumulation, and repeats the charge originally levelled at Ricardo by Malthus that the exceptions to which Ricardo's theory is subject may be encountered frequently (pp. xxxiv–xxxvi).

Even in his defence of Ricardo (1890) against the attacks of Jevons, Ingram, and the almost hysterical Adolph Held, Gonner's purpose is not so much so absolve Ricardo of either theoretical

error or of making extreme assumptions (though he does reject the still widely held misapprehension that Ricardo adopted a wages-fund theory), as it is to claim that the critics exaggerated their case in the cause of polemics. His defence of Ricardo against the strictures of Jevons, for example, seems to consist not in arguing that Ricardo's economics was on the correct lines, but rather that Ricardo himself could not have 'shunted the car of economic science onto the wrong lines' because he was no more than following a tradition established by others before him. It is also interesting to note that in this article, Gonner issues the claim (using it as a defence of Ricardo) that Ricardo never intended his *Principles* to be a complete coverage of the subject. This, of course, was the famous excuse for Marshall's wholesale re-interpretation of Ricardo, which has since been deprived of its veracity with the discovery and publication of Mill's correspondence with Ricardo over the book.

So far as can be determined, Gonner's life was uneventful. After graduating from Oxford, he took up his first academic post at Bristol in 1885. From there he moved in 1888 to the newly founded University College of Liverpool, eventually occupying (in 1891) the Brunner Chair of Economic Science in that College's institutional successor, the University of Liverpool, a post in which he remained until his death on 25 February 1922. Like so many British academics, he was seconded to government service during World War I, serving in the Ministry of Food – first as economic adviser and later as its Director of Statistics. From time to time he served as official arbitrator in industrial disputes for the Ministry of Labour.

Selected Works

1890. Ricardo and his critics. *Quarterly Journal of Economics* 4: 276–90.
1891. In *Principles of political economy and taxation*, ed. D. Ricardo. London: G. Bell & Sons.
1893. The survival of domestic industries. *Economic Journal* 3: 23–32.
1895. *The socialist state: Its nature, aims and conditions*. London: W. Scott.
1899. *The social philosophy of rodbertus*. London: Macmillan.
1906. *Interest and saving*. London: Macmillan.
1912a. *Common land and inclosure*. London: Macmillan.
1912b. The economic history. In *Germany in the nineteenth century*, ed. J.H. Rose, C.H. Herford, E.C.K. Gonner, and M.E. Sadler. Manchester: Manchester University Press.
1923. In *Economic essays*, ed. D. Ricardo. London: G. Bell & Sons.

Goods and Commodities

Murray Milgate

Keywords
Characteristics; Commodities; Demand theory; Exchange value; Goods and commodities; Jevons, W. S.; Johnson, H. G.; Lancaster, K.; Marginal revolution; Marshall, A.; Marx, K. H.; Menger, C.; Mill, J. S.; Produced and scarce commodities; Ricardo, D.; Robbins, L. C.; Scarcity; Separable utility functions; Smith, A.; Subjective theory of value; Water–diamonds paradox

JEL Classifications
A1

Towards the end of the 1950s, Harry Johnson produced the following theorem on value theory:

> Define a good as an object or service of which the consumer would choose to have more. Then the collection of goods he chooses when he has more money to spend (prices being constant) must represent more goods than that he chooses when he has less money to spend (since he could have had more of each separate good).

i. If his income rises, he buys more goods; this implies a presumption that normally the income effect is positive.

> ii. If he chooses collection B when he could have had collection A for the same money (i.e. $\sum p_b q_b = \sum p_b q_a$), he does not choose A if he could have had B for less money, because that would mean collection B represented less goods than collection A, and conflict with the definition of goods. Hence, when A is chosen B must be at least as expensive (i.e. $\sum p_a q_b \geq \sum p_a q_a$). This establishes that the substitution effect is non- negative (by subtraction, $\sum (p_b - p_a)(q_b - q_a) \leq 0$).
>
> Hence we derive both parts of the law of demand from the definition of goods. The hypothesis from which we have deduced it is that goods are goods. (1958, p. 149)

The idea that definition of goods carried with it the whole of the theory of demand, that the explanation of the determination of the exchangeable value of 'things' was intimately bound up with the definition of the 'things' themselves, probably struck many of Harry Johnson's readers as amusing. Indeed, Johnson may even have had this end in mind – no doubt many in the profession were beginning to wonder whether the frequency with which reconsiderations of something apparently so obvious as the theory of demand were being undertaken was entirely necessary. However, it did not strike at least one of his readers as being just another amusing aphorism: Kelvin Lancaster, upon whose review of Hicks's *Revision of Demand Theory* Johnson was commenting when he produced his theorem, took it seriously. Eight years later in the *American Economic Review* Lancaster advanced the so-called characteristics theory of demand. The argument was a simple corollary of the Johnson theorem: if it is the aim of the theory of demand to determine the prices of goods, then one ought to specify as clearly as possible the goods which are being demanded. After all, on this line of reasoning one demands not just physical objects, but the qualities with which they are endowed; it is to their *characteristics* that the potential purchaser first turns his attention.

The interesting features of this little episode, however, are not exhausted in a consideration of the ideas to which it gave rise. Quite as important are the implications which follow upon the recognition of the fact that the kinds of questions which lie behind Johnson's theorem had been debated before in contexts where certain useful results were generated. At least since the time of Adam Smith, economists have struggled to be clear about what it is in the nature of the things which are daily exchanged on markets that gives rise to exchangeable value. When Smith discussed the famous water–diamonds paradox, and drew from it (however perilously) the conclusion that the theory of exchangeable value should focus upon what may be called the objective conditions of production of things, rather than upon the subjective conditions of their consumption, he was engaged in just such an endeavour.

Smith was followed in this project by Ricardo. In the opening passages of the *Principles*, by establishing a clear line of demarcation between scarce and reproducible commodities, Ricardo reached Smith's conclusion by a different route. Marx praised this passage from Ricardo and focused his attention exclusively upon what he termed the commodity form. Moreover, this was not exclusively a classical preoccupation. Later writers, to whom modern economics seems to owe much more, also took the question very seriously indeed. Having returned to Smith's original paradox, they applied the distinction between total and marginal utility and to their satisfaction resolved it. This deprived Smith's original conclusion of its validity and allowed neoclassical writers to rebuild the theory of exchangeable value upon the basis of the subjective conditions of consumption of goods. Marshall was very clear about this at the beginning of the second chapter of Book II of his *Principles*.

The questions that Johnson's theorem prompts, therefore, include also those which were raised in these widely publicized and not insignificant debates of the 19th century over the distinction in the theory of value between those physical objects whose main characteristic is that they can be said to be in short supply, and those whose quantity may be increased by reproduction on an extended scale. To what extent, if at all, the choice of terminology by earlier writers reflects these differences is the subject matter of an investigation into goods and commodities.

Etymological Preliminaries

In English the word *good* derives from the Old English word *god*. It is related also to the Old Frisian *god*, the Old High German *guot*, the Old Saxon *god*, and the Old Norse *gòdr*. The word is defined in the *Oxford Dictionary of English Etymology* as 'the most general adjective of commendation'. The substantive plural form, *goods*, while sharing the origins, seems not to have appeared in English until the 13th century with a meaning much as it has today: objects or things which confer some advantage or produce some desirable effect upon their owner. Two further points may be noted. The first is that although there exists a genitive singular in Old Norse, no Teutonic language seems to have possessed a substantive plural form. Its usage in this manner probably derives from the Latin *bona*. The second is that despite the standard *O.E.D*. classification of *goods* as indeclinable, a substantive singular form has become common among economists.

In both modern French and German, the adjectives *bien* and *gut* share the meaning and sense as *good* in English (the German sharing the same Old Teutonic origins). The substantive plurals *biens* and *Güter* likewise share meaning and sense, together with the partial Latin origin of the English.

James Bonar's definition of the term *goods* in the original edition of this *Dictionary* – that 'by the plural (Goods) is denoted concrete embodiments of usefulness' – suggests that nothing of substance had been altered in the definition of the word even after it had been co-opted into the formal terminology of economic theory. Furthermore, Bonar's statement that *goods* are the physical embodiment of the metaphysical quality of *good*, seems to apply across all three languages. Of course, given that a substantive singular form is now in common usage, that part of Bonar's definition which went on to argue that the substantive singular *commodity* 'is employed by economists to represent the missing singular of goods', must be abandoned.

The word *commodity* is of entirely different origin and meaning. Its roots are in Latin and it is defined in the *Oxford English Dictionary* as 'a thing produced for use or sale, an article of commerce, an object of trade'.

Commodities

Questions as to the essential properties of the things exchanged in a market economy, though they had arisen in the work of earlier economists, took on an entirely new dimension with the commencement of the systematic study of exchangeable value in the last half of the 18th century. Following immediately upon the definition of wealth as the 'annual production' of the system, and the analysis of the effects of progress in the division of labour on the 'proportion between annual production and consumption', Adam Smith had confronted the issue of the valuation of this 'quantity of commodities annually circulated'. The problem was to establish, in the first instance, the sphere within which exchangeable value was to be examined. His answer, though failing to take into account conditions of relative scarcity, illustrates just as clearly as Johnson's theorem how an apparently neutral choice of language may be the bearer of certain theoretical precepts upon which an entire argument rests.

Perhaps even more importantly, Smith seems to have established not only the formal framework for the theory of value, but also the very language in which it was transmitted in orthodox circles right down to the time of Ricardo. That argument is sufficiently familiar not to have to be rehearsed here – the essential ingredient that is relevant to us is its rejection of the notion that the 'utility of some particular object' has anything to do with determining exchangeable value and that, instead, the exchangeable value of an object is to be explained in terms of what Smith variously called the 'toil and trouble of obtaining it' or its 'difficulty and facility of production'.

These objects are quite consistently called by Smith *commodities* and not *goods*. The terminology and the theoretical construct seem to match quite well. If one is to follow Smith into an investigation of the relationship between conditions of

production and relative prices, the usage of the term *goods* would be less than apposite. This, of course, is not to say that the familiar word *goods* does not crop up from time to time in the *Wealth of Nations* (opening the book at random would quickly disprove such a strong assertion). Nor is it to claim that Smith even bothered to take the time to explain his pattern of usage. But a determinate pattern there surely is.

Consider, for example, the discussion of the water–diamonds paradox, a passage where *goods* appears twice. This very short passage is followed by a carefully constructed paragraph setting out in a quite formal and purposeful way the project for the remainder of Book One. In that particular place, the term *commodities* is used exclusively. Indeed, the following three chapters, on real and nominal price, the component parts of price, and natural and market price, adhere fairly rigidly to this pattern of formal usage. The index, which was added to the original in its third edition of 1784, contains a lengthy entry under *commodities* but not one for *goods*. What is also interesting is that as between the two words, *goods* usually appears in those more discursive passages of the *Wealth of Nations*, whereas *commodities* is reserved for passages of a more formal, theoretical kind.

A remarkable parallel is to be found in the third edition of Ricardo's *Principles*. There, in the first paragraphs of the chapter on value, Ricardo makes a significant attempt to define just what it is that is important in the nature of those objects whose prices are determined on markets. At the same time, it should be noted, he replaces Smith's argument as to why exchangeable value is to be investigated in the sphere of production. The argument is pure Ricardo. Utility is 'essential to exchangeable value', objects which contribute in no way towards 'gratification' would be 'destitute of exchangeable value', but it does not determine it. Two conditions then remain to determine exchangeable value – Smith's difficulty and facility of production and, what Smith had passed over, scarcity. There follows Ricardo's famous twofold classification of commodities: those which are currently reproduced (produced commodities) and those which are fixed in quantity (scarce commodities). The exchangeable value of the former, when competition operates without restraint, is to be investigated in terms of the available methods of production. Relative prices of the latter depend upon the 'wealth and inclinations of those who are desirous of possessing them'. Ricardo restricts the investigation to produced commodities and his use of terms resembles the pattern one discerns in the *Wealth of Nations*.

This particular argument was taken up subsequently by two writers who stand in contrasting positions with respect to this classical conception of the framework for the analysis of exchangeable value – John Stuart Mill and Karl Marx. Both built quite self-consciously on the work of Smith and Ricardo. But as it happens, while Mill was effectively to put in place ideas (which admittedly had been in the air for some time) that were quite dramatically to modify the classical position, Marx was to revivify it. How closely terminological conventions reflect these factors is a question of some importance in the present context.

To begin with, let us turn to the German language, and to Marx, who claimed that Ricardo's argument for restricting the domain of the theory of value and distribution to the sphere of producedcommodities had been 'formulated and expounded in the clearest possible manner' (1859, p. 60). Quite unlike Ricardo, Marx not only consistently avoided the use of term *Güter* (*goods*) – it is hardly possible to forget that the first chapter of *Capital* bears the title *Wares* – but actually considered the theoretical consequences of these terminological conventions in the *Contribution to the Critique of Political Economy.* Though not especially satisfying in itself, the theoretical argumentation of the *Critique* is simple enough: 'use-value as such, since it is independent of the determinate economic form, lies outside the sphere of political economy' (1859, p. 28).

Goods

As the basis of the theory of value shifted away from the old classical idea of production as a circular process, towards the newer and different idea of an economic process resembling a one-

way street – from 'factors of production' to 'goods' – there began simultaneously a retreat from the examination of exchangeable value in terms of the objective conditions surrounding the production of commodities, and an advance towards a theory of value grounded in the subjective conditions surrounding the consumption of goods. Of course, this was not entirely an unprecedented idea (the work of Lauderdale and Bailey comes to mind), what is different is the fact that these notions are now placed on a firmer theoretical footing than had hitherto been the case and that they come to form the mainstream of the discipline.

The orientation thereby imparted to the theory of exchangeable value by the economists in the vanguard of this change took as its starting point precisely those passages of the *Wealth of Nations* which had been so important in establishing the conceptual apparatus of the earlier classical economists. But the lesson that was drawn from them was not that which had been drawn by Smith. They, too, were keenly interested in the properties of the actual objects of market exchange, but from Smith's water–diamonds paradox they did not reach the classical conclusion, but rather one that held that the joint conditions of scarcity and utility would act to determine relative prices. As Pareto was eloquently to put it, economics became the study of equilibrium between man's tastes and the obstacles to satisfying them. Exchangeable value, to borrow Jevons's terminology, would be determined by the final degree of utility.

The 1870s were, of course, the years in which the basic provisions of the new constitution of economics were laid down almost simultaneously in Britain, France and Germany. Precursors had been sought out and honoured by those in the vanguard of the new theory, and the battle against the 'noxious influence of authority', as Jevons put it, already promised success – even the sterner opposition of the historical school was beginning to seem less formidable. Yet despite these quite rapid developments, in the initial years of the marginal revolution the language and usage in English economics seems to have remained essentially as it had been in the classical period. An example may serve to highlight the point.

William Stanley Jevons, who by the second edition of his *Theory of Political Economy* in 1879 had succeeded in substituting 'economics' for the older term 'political economy' in everything but the title of his book, retained the substantive *commodities* even though it was to their want-satisfying qualities that he wished to defer in his explanation of exchangeable value. Usage of the term *goods*, which conveys with greater accuracy the theoretical conceptions at the base of this new approach to the theory of value, appears to have been consciously avoided by Jevons. There is a particularly interesting passage from the *Theory of Political Economy* that illustrates the degree to which Jevons grappled with the language in which to express his theory:

> It will be allowable … to appropriate the good English word *discommodity*, to signify any substance or action which is the opposite of *commodity*, that is to say, *anything which we desire to get rid of* … Discommodity is, indeed, properly an abstract form signifying inconvenience or disadvantage. (1871, p. 114, italics in original)

It is impossible to resist the temptation to add that 'the good English word' *discommodity* is of Latin origin, and that the formal introduction of the simple Old English word *goods* at this juncture would have relieved Jevons of the need to conduct such linguistic exercises (see also, Jevons 1882, p. 11). Nevertheless, the example is sufficient to indicate that in the English language at least, *goods* was not at this time in formal use in theoretical economics.

Despite this, however, the appearance of the term goods in the formal literature of economics is inextricably linked with the rise of the neoclassical theory of exchange and demand. But to see how this is so, it is necessary to turn to the writings of German economists of the new school.

In the German neoclassical literature, quite precise definitions were given for the formal usage of the substantives *Gut* and *Güter*. Carl Menger's *Grundsätze der Volkswirtschaftslehre* (1871) provides a particularly striking example of this:

> Diejenigen Dinge, welche die Tauglichkeit haben in Causalzusammenhang mit der Befriedigung menschlicher Bedürfnisse gesetzt zu werden,

> nennen wir Nützlichkeiten, wofern wir diesen Causalzusammenhang aber erkennen und es zugleich in unserer Macht haben, die in Rede stehenden Dinge zur Befriedigung unserer Bedürfnisse tatsächlich heranzuziehen, nennen wir sie Güter. (1871, pp. 1–2)

Menger, in fact, went so far as to devote an exceedingly long footnote (printed as an appendix to the *Werke* edition of the *Grundsätze*) to the history of the usage of this term in this sense.

How and when the equivalent term entered the formal language of English economics – and when it might be said to have established itself – is not a difficult question to answer. Alfred Marshall's *Principles* (1890) seems to be the innovator.

In a passage from the second edition of the *Principles* dating from 1891, Marshall remarked that lacking any short term in common use to represent all desirable things, that is 'things that satisfy human wants', he proposed 'to use the term *Goods* for that purpose' (1961, I, p. 54, italics in original). In the second edition Marshall appended a footnote to the effect that he intended to replace the singular *commodity* with the term *good*, and gives as explicit justification for this the correspondence between his usage and that of the German economists (see 1961, II, 185e). This appears to be the first systematic application of the term *goods* in the formal terminology of economic theory – what is more, its usage is derived from the German. Note that a substantive singular form also appears.

It would seem reasonable to conclude, therefore, that it is from this source (that is, from Marshall's *Principles*) that the term *goods* gained wide circulation in economic theory. The date by which it might be reckoned to have established itself would appear to be around the mid-1890s, as it was in the fourth edition of the *Principles* in 1898 that Marshall chose to delete the footnote alluded to in the previous paragraph. This would accord broadly with the date at which the original edition of this *Dictionary* appeared containing James Bonar's entry under the heading *goods*. It is interesting to note that by the end of the 1920s the term had been so fully absorbed into the language of economic theory that Robbins chose to omit from the English edition of Wicksell's *Lectures* a paragraph where the question of this terminology is discussed. According to Robbins's editorial note, this paragraph was 'of no interest to English readers' (Wicksell 1934, vol. 1, p. 15, n. l).

The originators of the modern theory of exchange and demand, nevertheless, had established a terminology through which to convey one of the basic tenets of their argument. Not only did they appreciate that goods are goods, but they expended a considerable amount of time and energy establishing that the subjective conditions of consumption were the appropriate place to locate the analysis of exchangeable value.

Characteristics

This brings us back to Harry Johnson's theorem. The grounding of a theory of exchangeable value upon the notion of *goods* was taken in certain to require a closer specification of the want-satisfying qualities of the 'things' which are daily exchanged on markets – since these are, in the final analysis, the *goods* which form the subject of the examination. When one contemplates the kinds of developments in the theory of exchange which might contribute towards the fulfilment of this requirement, nearly all of them seem to entail a widening of the gap between the actual 'things' which are exchanged on markets, and their want-satisfying characteristics which are the real subjects of demand. This, of course, is the direction in which the characteristics theory of demand has already taken us. Its problematic, of course, is to establish a transformation from characteristics to the actual objects through which these characteristics are transmitted. In the language of Lancaster, what is required is a well-defined mapping from the characteristics space to the goods space – since in the end the prices that are thrown up on markets are attached to actual objects and not to their characteristics. The theory of separable utility functions has been of immense assistance in this regard.

However, if the classification of these characteristics could be rendered sufficiently fine, then

a concomitant implication would seem to be that the idea of securing a theory of the prices attached to actual objects exchanged on markets would need to be sought in some other direction. Otherwise, we should be left with a theory of exchange and demand which made no contact, even at an abstract theoretical level, with the material realities of market exchange in modern economies. We should certainly have a theory of *goods*, but to what form of economic organization such a theory could be held to apply, if any, is not at all obvious.

So that such speculations should not be thought to be idle, it is interesting to note that language is not only a vehicle for the transmission of theoretical conceptions; it is often the vehicle through which a whole array of structural and cultural data about social interaction is conveyed. Exchange in different kinds of societies frequently embodies these complex social relations – so much so that the familiar idea of economists of the modern school that a universally applicable analysis of exchange is somehow desirable, or even possible, would seem to be fraught with pitfalls.

See Also

► Exchange

Bibliography

Jevons, W.S. 1871. *The theory of political economy*, ed. R. D.C. Black. Harmondsworth: Penguin, 1970.

Jevons, W.S. 1882. *The principles of economics*, ed. H. Higgs. New York: Kelley, 1965.

Johnson, H. 1958. Demand theory further revisited or goods are goods. *Economica* 25: 149.

Lancaster, K.J. 1966. Change and innovation in the technology of consumption. *American Economic Review, Papers and Proceedings* 56 (1–2): 14–23.

Maitland, J. (Earl of Lauderdale). 1804. *An inquiry into the nature and origin of public wealth*. New York: Kelley, 1962.

Malthus, T.R. 1827. *Definitions in political economy*. New York: Kelley, 1963.

Marshall, A. 1961. *Principles of economics*. 9th variorum edition. London: Macmillan. (1st edn, 1890.)

Marx, K. 1857. *Grundrisse*. Harmondsworth: Penguin, 1973.

Marx, K. 1859. *A contribution to the critique of political economy*. London: Lawrence & Wishart, 1971.

Marx, K. 1867. *Capital*. 4th edn, 1909. 3 vols, New York: International Publishers, 1967.

Mauss, M. 1925. *The gift*, ed. E.E. Evans-Pritchard. London: Routledge & Kegan Paul, 1970.

Menger, C. 1871. *Grundsätze der Volkswirtschaftslehre*. In *The collected works of Carl Menger*. Vol. 1. London: LSE. Reprints No. 17, 1934.

Mill, J.S. 1848. *Principles of political economy*. Peoples edn. London: Longmans, 1873.

Palgrave, R.H.I., ed. 1894–9. *Dictionary of political economy*, 3 vols. London: Macmillan.

Roscher, W. 1854. *Principles of political economy*. Trans. from the 13th German edn. Chicago: Callaghan and Coy, 1882.

Say, J.-B. 1803. *A treatise on political economy*. Trans. C.R.C. Prinsep. New York: Kelley. Reprinted, 1971.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*, 2 vols, ed. E. Cannan. London: Metheun, 1961.

Sraffa, P. (ed. with the collaboration of M.H. Dobb) 1951–73. *The works and correspondence of David Ricardo*. Cambridge: Cambridge University Press.

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

Wicksell, K. 1934. *Lectures on political economy*, 2 vols, ed. L. Robbins. London: Routledge & Kegan Paul.

Google

Hal R. Varian

Abstract

This article provides an overview of Google, focusing on its economic history, specifically on three topics of interest: the ad auction, the IPO auction and the role of economics at Google.

Keywords

Ad auctions; Auctions; Google; Internet economics; Search engines; Pagerank

JEL Classification

L86; D83

Introduction

Google is an Internet search engine founded by Sergey Brin and Larry Page in 1998. Its creation was stimulated by the NSF research program on digital libraries, which also led to the creation of the search engine Inktomi (NSF 2004).

Google was the first search engine to use the link structure of the web as an aid to determining relevance. The link structure determined relevance in two ways: the relevance of a document to a query depended on (1) the number and quality of web sites linking to it and (2) the relevance of the anchor text in the link to the query.

Exploiting these two factors required crawling the web and recording the link structure. The crawling in turn depended on developing computer infrastructure to store and access the massive and ever growing amounts of data. As part of this effort, Google had to invent data manipulation tools that enabled the storage, retrieval and manipulation of vast amounts of data.

The algorithm used to score web sites is known as 'pagerank'. The pagerank of a web page can be thought of as the probability that a person randomly clicking on links will arrive at that page. It can be calculated by a recursive procedure defined on the graph describing the link structure of the web.

Although the use of the link structure of the web was a significant improvement on search engines of the time, virtually every search engine now uses variants on this algorithm and attention has shifted to other indicators of relevance.

Company History

The domain name google.com was registered in September 1997. The company was incorporated a year later in September 1998 and Page and Brin moved operations from Stanford University to a wing of a residential house in Menlo Park. Contrary to popular accounts, the company did not operate out of the garage but was located in a separate wing of the house containing three small bedrooms. This wing was entered via the garage, which was actually used for storage and overflow.

In March 1999 Google moved to University Avenue in Palo Alto and shortly after that to an office complex in Mountain View. The original Google buildings were called 0 and 1. Later on they added two new buildings called π and e.

Shortly before the IPO in 2004, Google moved to a group of buildings at 1600 Amphitheatre Parkway in Mountain View that was previously occupied by Silicon Graphics. This became the centre of the so-called Googleplex. As of Spring 2012, Google occupied 45 buildings in Mountain View and had engineering and sales offices in many other locations around the world.

Auletta (2010), Battelle (2006), Levy (2009, 2011), Varian (2012) and others have written historical accounts of Google. This essay focuses specifically on three topics of particular interest to economics, namely the ad auction, the IPO auction and the role of economics at Google.

Google Ad auction

Google's business model is built around advertising. The original business model involved selling text ads located above the search results. The advertiser paid a fixed amount to have its ad displayed when a user query matched a particular keyword chosen by the advertiser. This is known as a 'pay per impression' pricing model.

The impression prices were set via the discretion of the sales force, but Google soon realised that this model did not scale to millions of keywords.

The idea of auctioning off ad space seems to have originated with GoTo, a company that came out of Bill Gross's IdeaLab. GoTo's original concept was to auction search result positions to the highest bidder using a first-price auction. This was not successful, so they tried selling ads (which they termed 'paid introductions') via a similar auction. This developed into a successful business which was later renamed Overture and subsequently acquired by Yahoo.

Customers were originally cautious about paying simply for ad positions since they didn't know how much they were worth. This prompted GoTo

to switch to a 'pay per click' model, which only charged the advertisers when someone clicked on their ad and navigated to their web page.

Google recognised the utility of using an auction model for selling ads and improved on the GoTo model in two ways. First, they priced the ads based on the estimated impression value, and second, they used a second-bid pricing model as opposed to the first-bid model used by GoTo.

The logic of the estimated impression value pricing was to maintain consistency with the pay per impression model Google originally used. The logic is based on the identity

$$\text{bid per impression} = \text{bid per click} \times \text{clicks per impression}$$

The advertiser would bid per click, Google would estimate the clicks per impression, and then the bid per click would be calculated and used to rank the ads and determine the price the advertiser paid.

The second innovation involved setting the price of the click to be the minimum price necessary to maintain the advertiser's position, a pricing rule now known as a 'generalised second price auction'. The rationale was that in a first-price auction an advertiser who wanted to minimise the price per click would experiment with lowering its price until it found the minimum amount necessary to maintain its desired position. This tended to make bidding somewhat unstable, so Google decided to just set the bid equal to the minimum required amount immediately.

To express this more formerly, let p_s be the price per click of the ad in position s, let e_s be the estimated clickthrough rate of the ad in position s, and b_s be the bid of the ad in position s. Then the estimated price per impression was

$$p_s e_s = b_{s+1} e_{s+1}$$

so the price per click became

$$p_s = b_{s+1} e_{s+1} / e_s$$

The Google ad auction was developed by Salar Kamanger and Eric Veach in the Fall of 2001 and went live in February 2002. A detailed account of its development is described in Levy (2009). More technical account are available in Edelman et al. (2007), Varian (2006, 2009) and Lahaie et al. (2007).

In May 2002 Google CEO Eric Schmidt invited Hal Varian to become a consultant at Google. Schmidt suggested that Varian 'take a look' at the ad auction since, he said, 'I think it might make us a little money'.

Varian examined the ad auction using the tools of auction theory and examined the structure of the Nash equilibria for the model (Varian, 2006). He presented this work at seminars at Stanford and the Federal Trade Commission, but did not seek to publish it as it was considered proprietary consulting work for Google.

In 2005 Edelman, Odlovsky and Schwartz independently developed a similar model of the Google auction. Varian asked for permission to publish his article and both the Edelman–Odlovsky–Schwartz and Varian papers appeared in 2006–07. Dozens of papers have subsequently been published that examine various aspects of the Google auction; see Lahaie et al. (2007) for a review.

Estimating Clickthrough Rates

The Google pricing model required estimation of clickthrough rates. The basic model was that the observed clicks per impression for ad a in position s was given by the product of an ad-specific effect x_a and a position-specific effect z_s. The position-specific effect for position s was determined by looking at the clicks received by random 'a's in position s and the ad-specific effect was determined by looking at the historical clickthrough rate of ad a, after normalising for the position where it occurred.

One problem with this system is that one had to serve many ad impressions in order to estimate the clickthrough rate accurately. This led to the development of a logistic regression model where the probability of a click depended on a number of other predictors in addition to historical performance. This logistic regression involved billions

of observations and so required specialised computation to estimate.

This system has subsequently evolved in a variety of directions, involving enhanced measures of 'ad quality', 'landing page quality' and more sophisticated ranking algorithms that take into account learning behaviour on the part of users.

IPO

In 2004 Google decided to make a public stock offering to raise a targeted amount of $2,718,281,828 (the digits coincide with the first few digits of e). Since the company's business model was built around an auction, Brin and Page decided to structure the IPO as an auction.

The idea was that each potential buyer could issue a bid per share and the number of shares desired at that price. The bids would be rank ordered and the price would be set at or near the point where the demand for shares equalled the supply. All shareholders who bid more than the offering price would receive the shares they ordered at the market-clearing price, or something close to it. (The market-clearing price was an input into the final pricing decision.)

The primary motivation for holding the IPO was to create liquidity for shareholder employees. It was felt that an auction structure would lead to a more stable aftermarket and avoid the volatility that had often accompanied high-tech IPOs. In addition, the auction was open to all investors, including small investors, rather than only 'qualified investors', as is typically the case with book-building IPOs.

Prior to this there had been several other examples of IPO auctions, most notably for privatisation of public utilities in Australia and Singapore, IPOs in Israel, and some small-scale experiments in the USA. The Google IPO was by far the largest and most prominent example of an IPO auction.

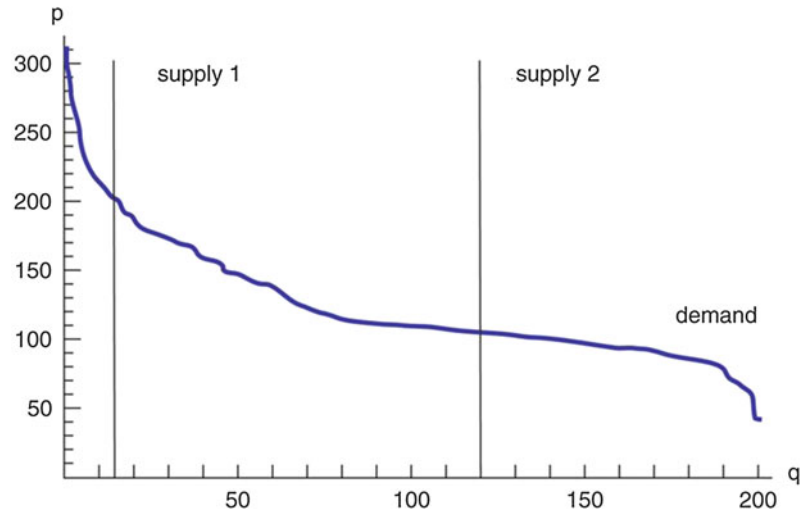
There were several interesting economic effects in the IPO auction that have not been recognised in the literature, which are briefly outlined here.

High Estimated Price

The S1 filing estimated that Google shares would sell in the range $108–$135. The pricing was deliberately chosen to be on the high side so as to appeal to investors who had a long-run viewpoint, and discourage short-term speculation.

Speculative Excess

It is likely that the divergence of opinion with respect to valuation is larger for amateur investors than professional investors. If the amateur investors have higher values on average than the professionals we would get a demand curve for shares of the shape shown in Fig. 1. The steep part of the



Google, Fig. 1 Hypothetical demand for a stock

curve is composed of the amateur investors, while the flat part is composed of professional investors. If the equilibrium occurs at the position indicated by 'supply 1', subsequent movements in the supply can lead to large changes in price while an equilibrium at position 'supply 2' would tend to have a more stable aftermarket.

Participation Incentives

In the book-building IPO model, investors tend to anticipate a first-day pop, which provides a reason for them to participate in the IPO rather than wait. If investors in the auction model believe that the auction will determine the market-clearing price, with no pop, the investors may as well just avoid the bother of participating in the auction and just buy on the opening day. Hence, if the bidders believe that the auction will accurately determine the market price, they have no reason to participate. But if participation is thin, the auction is unlikely to determine the market-clearing price.

DK Risk

The term DK risk refers to 'decay' of interest in the stock. In an IPO one generally buys the stock by placing an order with a broker, but only has to pay for the stock three business days later. If the stock price is higher on the third day than on opening day, it is easy to get investors to pay up. If it is lower, it can be quite hard to get investors to pay, since they would have to pay more for the stock than the price available in the open market. With professional investors, brokers can insist on getting paid, although they don't like to strong-arm their best customers. However, if the third day has a price lower than the opening price thousands of small investors could try to avoid paying for the stock, which would be a nightmare for the investment bankers. This is yet another reason why a first day (really a third day) pop is viewed as highly desirable by investor bankers underwriting the stock.

Cheap Talk

Those who commit to buying a stock in a book-building IPO will generally receive the stock at something close to the price they committed to. Since the price they pay is more or less fixed, it is in their interest to talk the stock up, so the aftermarket price is high. Compare this to the situation where investors commit to buying a stock in the auction model. In this case they will receive the stock at the market-clearing price as long as their bid exceeds that price. This means that is in their interest to talk the stock stock down to reduce the price they have to pay. Prior to the Google IPO there was quite a bit of negative chatter about Google, perhaps for this reason.

Despite the difficulties described above, the IPO auction used by Google worked reasonably well. The IPO price was $85 and the stock closed at $100 on the first day. Six months later, when the employee lock-down expired, the stock was selling for about $200, providing employee liquidity and rewarding the hard work of those who built the company. Subsequently the stock increased substantially, hitting a high of around $750 a few years later.

Economics at Google

After developing the theoretical model of the ad auction in 2002, I worked on forecasting query growth and revenue growth, advertiser lifetime value, the AdSense for Content auction, and auction simulator and a number of other topics as a consultant to Google.

By 2007 the company had grown substantially, which, not surprisingly, led to increased demand for economic analysis. In the summer of 2007 I decided to retire from UC Berkeley and join Google full time. I set about hiring a team of economists, statisticians, engineers, operating researchers and other 'quants' to address the various business issues that arose at Google. In addition to this group, there was a small group of computer scientists at Google Labs who worked on 'algorithmic mechanism design', which studies market-like algorithms such as the ad auction.

In 2012 we had an additional influx of economists recruited from Yahoo Labs. As of 2012 there were about 15 PhD-level individuals working in economics-related fields at Google. Their work involves revenue modelling, marketing analytics, public policy, legal analytics, auction design and strategy, and decision support.

Subsequently, other companies, such as Yahoo, Microsoft, Amazon, eBay and Facebook, have recognised the contributions that economists can make to tech companies and have also created teams of economists.

See Also

- ▶ Auctions (Experiments)
- ▶ Auctions (Applications)
- ▶ Electronic Commerce
- ▶ Internet and the Offline World
- ▶ Online Platforms, Economics of

Bibliography

Auletta, K. 2010. *Googled: The end of the world as we know it*. New York: Penguin.

Battelle, J. 2006. *The search: How google and its rivals rewrote the rules of business and transformed our culture*. New York: Portfolio Trade.

Edelman, B., M. Ostrovsky, and M. Schwartz. 2007. Internet advertising and the generalized second price auction. *American Economic Review* 97(1): 242.

Lahaie, S., D.M. Pennock, A. Saberi, and R.V. Vohra. 2007. Sponsored search auctions. In *Algorithmic game theory*, ed. N. Nisan, T. Roughgarden, E. Tardos, and V.V. Vazirani. Cambridge: Cambridge University Press. Ch. 28.

Levy, S. 2009. Secret of googlenomics: data-fueled recipe brews profitability. *Wired*, 17(6). Available at: http://www.wired.com/culture/culturereviews/magazine/17-06/nep_googlenomics?currentPage=all. Accessed 9 Feb 2013.

Levy, S. 2011. *In the plex: How Google thinks, works, and shapes our lives*. New York: Simon & Schuster.

NSF. 2004. *On the origins of Google*. Available at http://www.nsf.gov/discoveries/disc_summ.jsp?cntn_id=100660. Accessed 9 Feb 2013.

Varian, H.R. 2006. Position auctions. *International Journal of Industrial Organization* 24(7): 1–10. Available at: http://www.ischool.berkeley.edu/research/publications/varian/2007/position. Accessed 9 Feb 2013.

Varian, H.R. 2009. Online ad auctions. *American Economic Review* 99(2): 430–434. Available at: http://www.aeaweb.org/articles.php?doi=10.1257/aer.99.2.430. Accessed 9 Feb 2013.

Varian, H.R. 2012. Economics of internet search. In *Handbook on the economics of the internet*, ed. J.M. Bauer and M. Latzer. Basingstoke: Edward Elgar.

Gordon, Robert Aaron (1908–1978)

B. G. Hickman

Keywords

American Economic Association; Business cycles; Cowles Commission; Econometric modeling; Gordon, R. A.; National Bureau of Economic Research; Price theory

JEL Classifications

B31

Gordon was born 26 July 1908 in Washington, DC, and died 7 April 1978 in Berkeley, California. He was a policy-oriented economist whose research style was quantitative but not econometric, and whose influence was felt not only through his basic research but also as a dedicated and tireless public servant. After his formative years as graduate student and instructor at Harvard University, 1929–38, he accepted a position at the University of California (Berkeley), where he remained until his retirement from teaching in 1976.

His early work in industrial organization culminated in the influential volume on *Business Leadership in the Large Corporation* (1944), which is noteworthy for its pioneering use of empirical data in the field. Another major strand is his work on unemployment, its structural and cyclical causes, and the goal of full employment (1962, 1967).

His lifelong research interests were primarily focused, however, on business cycles and their

causes. He championed the quantitative-historical method of cycle analysis over the National Bureau of Economic Research (Burns-Mitchell) and econometric modelling (Cowles Commission) approaches (1949), and he devoted much of his career to implementing his approach in studies of the business fluctuations of the interwar period (1951) and before and after the Second World War (1955a, 1969, 1974). His eclectic analysis of the causes of business cycles emphasized the Schumpeter-Hansen distinction between major and minor business cycles (1956) and attributed the major cycles to the appearance and exhaustion of investment opportunities in particular industries (1955b).

He combined a formidable talent for economic analysis with a sense of historical and institutional relevance and was impatient with the tendency he discerned in the profession at large to favour rigour over relevance in economic theorizing. This impatience was expressed at an early stage with regard to price theory (1948), reiterated in his presidential address to the American Economic Association on 'Rigor and Relevance in a Changing Institutional Environment' (1976), and repeated in his last piece on 'A Skeptical Look at the "Natural Rate" Hypothesis' (1978).

His contributions to the public weal were many and lasting, but two in particular may be cited. In 1956–59 Gordon undertook a massive study of business education jointly with James E. Howell for the Ford Foundation, and their 1959 report provided the stimulus for a radical reorientation of MBA programmes in graduate business schools toward the use of analytical methods drawn from economics, statistics, and the behavioural sciences. He also served as chair of the President's Committee to Appraise Statistics on Employment and Unemployment in 1961–62, which led to important reforms in the nation's statistics in this vital area.

Selected Works

1944. *Business leadership in the large corporation.* Washington, DC: Brookings Institution.

1948. Short-period price determination in theory and practice. *American Economic Review* 38, 265–288.

1949. Business cycles in the interwar period: The 'quantitative-historical' approach. *American Economic Association, Papers and Proceedings* 39(May): 47–63.

1951. Cyclical experience in the interwar period: The investment boom of the twenties. In Universities-National Bureau Committee, *Conference on business cycles*. New York: NBER.

1955a. Investment opportunities in the US before and after World War II. In *The business cycle in the post-war world*, ed. Erik Lundberg. New York: Macmillan.

1955b. Investment behavior and business cycles. *Review of Economics and Statistics* 37: 23–34.

1956. Types of depression and programs to combat them. In Universities–National Bureau Committee, *Policies to combat depression.* Princeton: Princeton University Press.

1959. (With James E. Howell.) *Higher education for business.* New York: Columbia University Press.

1962. (Co-author.) *Measuring employment and unemployment.* Report of the President's Committee on Employment and Unemployment Statistics. Washington, DC: Government Printing Office.

1967. *The goal of full employment*. New York: Wiley.

1969. The stability of the U.S. economy. In *Is the Business cycle obsolete*? ed. Martin Bronfenbrenner. New York: Wiley.

1974. *Economic growth and instability: The American record.* New York: Harper.

1976. Rigor and relevance in a changing institutional setting. *American Economic Review* 66: 1–14.

1978. A skeptical look at the 'natural rate' hypothesis. In *Economic theory for economic efficiency: Essays in honor of Abba P Lerner*. Boston: MIT Press.

Gorman, W.M. (Terence)

Charles Blackorby, Daniel Primont and
R. Robert Russell

Abstract

W.M. (Terence) Gorman was a theorist's theorist who nonetheless thought himself to be very practical. His contributions were brilliantly original – both in conception and in technical execution – but they often were a difficult read. We sketch below two of his lifelong interests – aggregation over agents and aggregation over commodities – and discuss briefly a few of his other contributions. We hope this brief outline will encourage others to peruse his works in more detail.

Keywords

Aggregation over agents; Aggregation over commodities; Cambridge School; Capital aggregation; Characteristics; Closed form representation; Compensation principle; Concavity; Diminishing marginal utility; Duality; Econometric Society; Engel curve; Equivalence scales; Expected utility theory; Expenditure function; General equilibrium; Gorman, W. M.; Hotelling's lemma; Indirect utility functions; Interpersonal utility comparisons; Le Chatelier principle; Marginal product of capital; Price aggregation; Profit functions; Roy's theorem; Separability; Two-stage budgeting

JEL Classifications

B31

Introduction

W.M. Gorman lived from 1923 to 2003. He graduated from Trinity College, Dublin, in 1948 in economics and a year later in mathematics. He held posts at the University of Birmingham, Oxford University, and the London School of Economics. Among many honours, he was the president of the Econometric Society in 1972, a fellow of the British Academy, and an honorary foreign member of the American Academy of Arts and Sciences and of the American Economic Association. (For a more detailed discussion of Gorman's career and work, see Honohan and Neary 2003.)

Gorman was interested in economics interpreted very broadly. He read history, philosophy, mathematics, and statistics. Nevertheless, most of his work was very abstract and many readers found it difficult to follow. We begin here by discussing his long-standing interest in two kinds of aggregation, over agents and over commodities, before considering some of his other contributions.

Aggregation Across Agents

Gorman's first published paper (Gorman 1953) provided necessary and sufficient conditions for the existence of a representative consumer. That is, given that each consumer in society has well-behaved preferences and demand functions,

$$x^h = f^h(p, m_h) \text{ for } h = 1, \ldots, H, \tag{2.1}$$

when can aggregate demand, $\sum_h x^h$, be generated by the demands of an aggregate agent so that

$$x = \sum_h x^h = f(p, m), \tag{2.2}$$

where $m = \sum_h m_h$? He demonstrated that a necessary and sufficient condition is that consumers have affine parallel Engel curves,

$$x^h = a(p)m_h + b^h(p), \tag{2.3}$$

so that the demands of the aggregate consumer can be written as

$$x = a(p)m + b(p), \tag{2.4}$$

where $b(p) = \sum_h b^h(p)$. In a later paper, Gorman (1961) provided a closed form representation of the preferences of consumer with parallel affine Engle curves: the indirect utility functions can be written as

$$V^h(p, m_h) = A(p)m_h + B^h(p) \text{ for } h = 1, \ldots, H, \tag{2.5}$$

and the indirect utility function of the representative consumer can be written as

$$V(p, m) = A(p)m + B(p), \tag{2.6}$$

where $B(p) = \sum_h B^h(p)$. Using Roy's theorem, the reader can easily check that (2.3) and (2.4) are derived from (2.5) and (2.6) respectively. The intuition behind this result is straightforward: if we take a dollar away from consumer 1 and give it to consumer 2, aggregate demand cannot change, because total income has not changed; the two consumers therefore must have identical marginal propensities to consume. There have been many attempts to generalize this notion of a representative agent (see, for example, Muellbauer 1976) but all suggest that aggregate demand for the *i*th commodity can be written as

$$x_i = \sum_{j=1}^{J} a_{ij}(p)b_j(m, z), \tag{2.7}$$

where z is a vector of characteristics. Gorman (1981) demonstrates that utility maximization requires that the rank of the matrix with elements a_{ij} be less than or equal to 3. Note that the idea of a representative consumer given in (2.3) entails that the matrix have rank 2. This proved that there is not much more that could be done to generalize this idea.

Over the course of the 1960s a battle raged between the US Cambridge and the English Cambridge over – to put it crudely – whether the marginal product of capital was a meaningful aggregate concept (see Fisher and Monz 1992, for an extensive discussion of this issue). Gorman (1968a) approached this problem in the dual, using profit functions. His argument in favour of using the dual was that it was important to model problems in the appropriate variables. When thinking about capital aggregation he noted that the individual agents all face the same prices. Hence he thought the problem should be attacked using profit rather than production functions.

In the simplest terms, suppose that each firm has a profit function given by

$$\pi^f\left(p, k^f\right) = \max_{y^f} \left\{p \cdot y^f \,|\, y^f \in T^f\left(k^f\right)\right\} \text{ for } f = 1, \ldots, F, \tag{2.8}$$

where k^f is a vector of the quasi-fixed factors of firm f and $T^f(k^f)$ is the technology set of the firm for each fixed value of the k^f. In this context, the marginal product of capital only makes sense if there exists an aggregate profit function that can be written as

$$\pi(p, k) = \max_y \{p \cdot y | \, y \in T(k)\}, \tag{2.9}$$

where k is a scalar measure of the aggregate capital stock. Gorman showed that a necessary and sufficient condition for (2.9) is that

$$\pi^f\left(p, k^f\right) = \alpha(p)\varphi^f\left(k^f\right) + \beta^f(p) \text{ for } f = 1, \ldots, F, \tag{2.10}$$

so that the aggregate capital stock is given by

$$k = \sum_f \varphi^f\left(k^f\right), \tag{2.11}$$

thus demonstrating that the marginal product of capital is not likely to be a meaningful concept very often. (For the implied restrictions on the technology sets and production functions, see Blackorby and Schworm 1984.)

The intuition for this result is a little less straightforward. Think of $\varphi^f(k^f)$ as a measure of bolted-down capital of firm *f*. Then, $\alpha(p)$ is clearly the marginal product of capital in firm f and (2.10) requires this be the same in every firm so that the aggregate profit function can be written as

G

$$\pi(p,k) = \alpha(p)k + \beta(p) \qquad (2.12)$$

where k is given by (2.11) and $\beta(p) = \sum_f \beta^f(p)$.

Aggregation Across Commodities

Gorman's best-known paper in this area (1959a) was written in response to a paper by Strotz (1957) on two-stage budgeting. Suppose that a consumer (or any organization with a well-behaved objective function and an expenditure constraint) wants to simplify its purchasing decisions as follows: first it wants to allocate funds optimally to broad categories of commodities and then later make the detailed calculations of how to spend the funds within any particular category. For example, a consumer might first decide how much money to budget for food and then later decide exactly how to allocate the food budget to particular commodities. The latter decision turns out to be equivalent to the separability of the commodities in each category, whereas the former hinges on a notion of price aggregation. (For more on two-stage budgeting and separability, see separability.)

Separability seemed a natural assumption to Gorman; it allowed the researcher to focus on a particular group of commodities without having to worry very much about anything else. In Gorman (1968b), however, he showed that this could have previously unknown implications. Suppose that one assumed that two groups of commodities were separable from their complements so that the utility function could be written as

$$U(x) = \overline{U}\left(U^A\left(x^A\right), x^C\right) \text{ and } U(x) = \hat{U}\left(U^B\left(x^B\right), x^D\right). \qquad (3.1)$$

Suppose in addition that some of the commodities in group A are not in group B, some are in both A and B, and the others are in B but not in A. Gorman showed that this implied that the utility function could in fact be written as

$$U(x) = \overset{*}{U}\left(U^a(x^a) + U^b\left(x^b\right) + U^c(x^c)\right), \qquad (3.2)$$

where the variables in a are those that are in A and not in B, those in b are in B but not in A, and those in c are in both groups. This surprising result can be taken in two ways: (1) perhaps too much separability is a dangerous thing but (2) on the other hand this theorem provides a fundamental and clean way of modelling additivity.

Ever attuned to different ways of looking at things, Gorman (1970), in a paper published only in 1995, noted that if a quasi-concave utility function could be written additively, as would often be required to study intertemporal or uncertain events,

$$U(x) = \overset{*}{U}\left(\sum_{s=1}^{S} U^s(x^s)\right), \qquad (3.3)$$

then at most one of the functions U^s could fail to be concave. Thus, for example, if the U^s functions were the same as in expected utility theory (that is, linear transformations of one another), then quasi-concavity would imply concavity. Again this can be taken two ways – as an easy way to obtain concavity or as the danger of too much separability.

General Interests

Gorman was well known for having widely circulated and widely cited but unpublished papers. Perhaps, the best known of these was 'A Possible Procedure for Analysing Quality Differentials in the Egg Market', a 1956 working paper of the Iowa Agricultural Experiment Station. (This paper was eventually published as Gorman, 1980. A more complete discussion can be found in Honohan and Neary 2003.) Suppose that it is not the commodities that consumers want but rather the characteristics embodied in them. Thus type i egg would contain a_i of characteristic A, b_i of characteristic B, and so on. Further, Gorman assumed that if one bought x_1 of type 1 egg and x_2 of type 2, then the total amount of characteristic A would be $a_1x_1 + a_2x_2$. Thus, at arbitrary prices only two types of eggs would be purchased if there were only two characteristics that were relevant and three types if three characteristics were

relevant, except in the degenerate case where relative prices were just right. Then the consumer would be indifferent between two or three or more types of eggs. Gorman argued that equilibrium prices should not contain any arbitrage opportunities and hence the degenerate case would be the normal one. This then suggested an agenda for empirical work that he and students pursued over the years. (Not very much of this research actually surfaced, although a 1959 University of Birmingham working paper entitled 'Demand for Fish: an Application of Factor Analysis' and his 1972 presidential address to the Econometric Society, 'A Sketch for the Demand for Related Goods', were widely cited for some time.)

Although Gorman wrote and published little in welfare economics, what he did write demonstrated a profound understanding of the issues involved. In order to avoid the known problem with the Kaldor (1939) compensation principle (namely, that situation B could be preferred to situation A by the Kaldor compensation criterion and that A could be preferred to B by the same criterion), Scitovsky (1941) had proposed a new test that required that B should be preferred to A if B was preferred by the Kaldor compensation test and A was not preferred to B by the same test. In Gorman (1955) he demonstrates in a very elegant manner that the Scitovsky criterion is intransitive; that is, that B could be preferred to A by the Scitovsky criterion, C preferred to B by the same criterion, and then that A could be preferred to C by the same test.

In a rather more philosophical vein, Gorman (1959b) presents a series of arguments that might lead one to think that social indifference curves are convex. In this he was clearly aware of the problems of interpersonal comparisons of utility and the idea of diminishing marginal utility for each individual. Even now this makes for a thought-provoking read.

Although Gorman did very little research in general equilibrium, when preparing a paper for a Festschrift in honour of Ivor Pierce he felt that only a general equilibrium paper would be appropriate. Gorman (1984) prepared an analysis of the Le Chatelier principle in general equilibrium. To give the reader the flavour of his elegant argument, let $\Pi(p)$ and $\pi(p)$ be the long-run and short-run profit functions of a firm. It seems natural to assume that

$$\Pi(p) \geq \pi(p). \tag{4.1}$$

Suppose that at $p = \overline{p}$ we are at a long-run equilibrium, so that

$$\Pi(\overline{p}) = \pi(\overline{p}). \tag{4.2}$$

Thus, the price vector $p = \overline{p}$ minimizes the difference between long-run and short-run profits, $\Pi(p) - \pi(p)$. The second-order condition for this minimization problem is given by

$$\sum_i \sum_j \Pi_{ij}(\overline{p})\theta_i\theta_j \geq \sum_i \sum_j \pi_{ij}(\overline{p})\theta_i\theta_j. \tag{4.3}$$

By Hotelling's lemma, this implies immediately that

$$\frac{\partial \overline{X}_i}{\partial p_i} \geq \frac{\partial \overline{x}_i}{\partial p_i}; \tag{4.4}$$

that is, the long-run response of commodity i to a change in its price is greater than or equal to the short-run response – the Le Chatelier principle.

Technically, much of Gorman's work was difficult; he frequently employed transformations of variables and functions with such speed that the reader felt, if not lost, at least dizzy. In Gorman (1976) he wrote what he thought of as a reasonable introduction to some of these tools and called the paper 'Tricks with Utility Functions'. We conclude with a discussion of one of these tricks and the lesson that Gorman thought could be learned from it.

Gorman begins his discussion of equivalent adult scales (equivalence scales) with a quote from a former schoolmaster who said 'When you have a wife and a baby, a penny bun costs threepence'. Consider a family of type $a = (a_1, \ldots, a_n)$ whose utility function could be written – after Barten (1964) – as

$$u_a = U^a(x) = U(x^a), \tag{4.5}$$

where the adjusted consumption vector,

$$x^a = \left(\frac{x_1}{a_1}, \ldots, \frac{x_N}{a_N}\right), \tag{4.6}$$

corrects for the number of equivalent adults. Note that the second equal sign implies that all households have the same utility function but defined on the adjusted variables. The expenditure function dual to U^a in (4.5) is given by

$$\begin{aligned} E^a(u_a, p) &= \min_x \{p_1x_1 + p_2x_2 + \ldots + p_Nx_N | U^a(x) \geq u_a\} \\ &= \min_{x^a} \{a_1p_1x_1^a + \ldots + a_Np_Nx_N^a | U(x^a) \geq u_a\} = E(u_a, p^a), \end{aligned} \tag{4.7}$$

where $p^a = (a_1p_1, \ldots, a_Np_N)$ and E is the expenditure function dual to U in (4.5). Thus, for a family of three bread-equivalent adults 'a penny bun costs threepence'. The adjusted compensated demand for good i is

$$x_i^a = E_i(u_a, p^a), \tag{4.8}$$

so that the ordinary compensated demand is

$$x_i = a_iE_i(u_a, p^a). \tag{4.9}$$

From this, the compensated demand elasticities can be written as

$$\varepsilon_{ij} = \delta_{ij} + \alpha_{ij}, \tag{4.10}$$

where δ_{ij} is the Kronecker delta and

$$\alpha_{ij} = \frac{\partial \ln x_i^a}{\partial \ln a_j} \tag{4.11}$$

at a fixed u_a is the compensated elasticity with respect to family size. It is easy from here to calculate the uncompensated elasticities as well. From this Gorman concludes that 'Were the theory true, and were the sample to include a great enough variety of family types, we could use them to calculate the price elasticities from survey data. As long as everyone faces the same prices, we need not even know what they are'. If one had tried to do this analysis, as Barten did, working only with (4.5) the simplicity of this model would not have been exposed.

This paper had dozens of such 'tricks', techniques that use either duality or separability arguments, and we recommend them to the reader.

See Also

- ▶ Duality
- ▶ Engel Curve
- ▶ Equivalence Scales
- ▶ Indirect Utility Function
- ▶ Separability

Selected Works

1953. Community preference fields. *Econometrica* 51:63–80.
1955. The intransitivity of certain criteria used in welfare economics. *Oxford Economic Papers* 11: 23–55.
1959a. Separability and aggregation. *Econometrica* 27:469–481.
1959b. Are social indifference curves convex? *Quarterly Journal of Economics* 73:485–96.
1961. On a class of preference fields. *Metroeconomica* 13:53–6.
1968a. Measuring the quantities of fixed factors. In *Value, capital and growth: Papers in honour of Sir John Hicks.* Edinburgh: Edinburgh University Press.
1968b. The structure of utility functions. *Review of Economic Studies* 32:369–90.
1970. The concavity of additive utility functions. Manuscript, University of North Carolina. In Gorman (1995).
1976. Tricks with utility functions. In *Essays in Economic Analysis,* ed. M. Artis and R. Nobay. Cambridge: Cambridge University Press.
1980. The demand for related goods: a possible procedure for analysing quality differentials in the egg market. *Review of Economic Studies* 47:843–56.
1981. Some Engel curves. In *Essays in the Theory and Measurement of Demand in Honour of Sir*

Richard Stone, ed. A. Deaton. Cambridge: Cambridge University Press.
1984. Le Chatelier and general equilibrium. In *Demand, Equilibrium and Trade: Essays in Honour of Ivor F. Pearce,* ed. A. Ingham and A.M. Ulph. London: Macmillan.
1995. *The Collected Works of W.M. Gorman,* vol. 1, ed. C. Blackorby and A. Shorrocks Oxford: Oxford University Press.

Bibliography

Barten, A. 1964. Family composition, prices, and expenditure patterns. In *Econometric analysis for national economic planning*, ed. P.E. Hart, G. Mills, and J.K. Whitaker. London: Butterworths.
Blackorby, C., and B. Schworm. 1984. The structure of economies with aggregate measures of capital: A complete characterization. *Review of Economic Studies* 51: 633–650.
Fisher, F.M., and J. Monz. 1992. *Aggregate production functions and related topics*. Cambridge, MA: MIT Press.
Honohan, P., and J.P. Neary. 2003. W.M. Gorman. *Economic and Social Review* 34: 195–209.
Kaldor, N. 1939. Welfare propositions of economics and interpersonal comparisons of utility. *Economic Journal* 49: 549–552.
Muellbauer, J. 1976. Community preference fields and the representative consumer. *Econometrica* 44: 979–999.
Scitovsky, T. 1941. A note on welfare propositions in economics. *Review of Economic Studies* 9: 77–88.
Strotz, R. 1957. The empirical implications of a utility tree. *Econometrica* 25: 269–280.

Goschen, George Joachim, Viscount (1831–1907)

P. Bridel

British statesman and financier of German origin, born in London on 10 August 1831; died at Seacox Heath, Surrey, on 7 February 1907. Goschen joined his father's firm of merchant bankers in London on leaving Oxford University. He became a director of the Bank of England in 1858 and an MP in 1863. He was given his first cabinet appointment in 1866. As Chancellor of the Exchequer under Lord Salisbury, Goschen's brilliant political career is chiefly remembered for his conversion and consolidation of the greater part of the National Debt and his reform of the gold coinage. He also set up two important Royal Commissions – that on the Depression of Trade and Industry in 1886, and the Gold and Silver Commission of the following year. Alfred Marshall's written Memoranda and Oral Evidence for both Commissions remained for nearly 40 years the half-forgotten source from which grew the Cambridge 'oral tradition' in monetary theory.

Goschen did not claim to be a professional economist though he wrote a number of essays and addresses on economic and monetary subjects. His famous *Theory of the Foreign Exchanges* (1861) remained for decades the standard work of reference on the subject. Although rather new in its blend of theory and facts, Goschen's book is extremely traditional in its analysis of the mechanis whereby international price adjustments are brought about. Building on Mill's version of the Hume-Ricardo quantity theory-cum-specie-flow doctrine, Goschen offers a clear presentation of how a country's exchange rate is determined by the amount of its short term indebtedness, the size of its monetary stock and the domestic price level. His discussion of the working of the Gold Exchange Standard (in particular the gold points mechanism) and an explicit statement of the purchasing power parity theory still repay study. However, his free trade convictions and his convinced support for all *laissez faire* policies led him to some debatable propositions in his policy chapters. In particular, his strong desire to avoid all interference with what he called the 'natural' workings of the price system made him argue that the exchange rate market is self-adjusting and that, in altering its discount rate, the Central Bank is *following* rather than controlling market conditions. Bearing in mind the extensive use of the discount rate policy made at the time by the Bank of England this argument –neither grounded in facts nor in theory – gave rise to much debate and criticism in Goschen's own time.

In 1890 the Royal Economic Society was founded, Goschen holding the office of president from that year until his death.

Selected Works

1861. *The theory of foreign exchanges*. London: Effingham Wilson, 1890.
1905. *Essays and addresses on economic questions*. London.

Gossen, Hermann Heinrich (1810–1858)

Jürg Niehans

Keywords

Gossen, H. H.; Imputation; Intermediate products; Invisible hand; Libertarianism; Marginal rate of substitution; Marginal revolution; Marginal utility; Optimal allocation; Optimal saving; Rent; Time preference; Value theory

JEL Classifications

B31

Gossen was born in Düren (between Aachen and Cologne) on 7 September 1810; he died in Cologne on 13 February 1858. Little is known about his life, partly because the inconspicuous bachelor did not attract attention, partly because most of those who had known him were dead by the time he became famous, partly also because his literary remains, scant as they must have been, are lost. The principal biographical source is the essay by Walras (1885). The available facts are admirably surveyed by Georgescu-Roegen (1983), on whose masterly introduction to the English translation of Gossen's book the following life sketch is mostly based.

Gossen's father was a tax collector under Napoleon and subsequently the Prussian administration; later he managed his wife's estate near Godesberg. Hermann obtained a good high-school education, showing ability in 'elementary mathematics', but his mathematical training never went beyond that level. Since his father insisted on a government career in the tradition of his forebears, his university studies in Bonn and Berlin concentrated on law and government.

In 1834, Gossen entered the civil service as a 'Referendar' (junior law clerk) in Cologne. While he seems to have been a well-mannered young man, the performance of his duties left much to be desired. He simply had no interest in a government career and loved the good things in life. There were complaints and reprimands, and the promotion to the rank of 'Regierungsassessor' came rather later than usual. Finally, in 1847, though his superiors seem to have shown considerable sympathy, he had no choice but to resign.

The transition to a new career was perhaps eased by his father's death, which spared him recriminations about his failure and provided him with the means for a new start. Gossen went to Berlin, where he seems to have sympathized with the liberal revolution, and then returned to Cologne as a partner in a new accident insurance firm. He soon withdrew from the firm, but continued to devise grandiose insurance projects.

Living with his two sisters, Gossen now devoted most of his energies to developing the unorthodox ideas he had expressed in his civil service examination papers into his magnum opus. The preface suggests that he hoped this would not only make him the Copernicus of the social universe but also open the door to an academic career. In 1853 an attack of typhoid fever undermined his health, and the disappointment about the fate of his book depressed him. Death came from pulmonary tuberculosis. He seems to have been an amiable, sincere and idealistic human being with broad interests, including music and painting. Brought up a Catholic, he developed into an enthusiastic hedonist. Dreaming of reforming the world, he lacked the force to conquer it.

The *Entwickelung der Gesetze des menschlichen Verkehrs* was published in 1854 at Gossen's expense by the publisher Vieweg in Brunswick. Very few copies were sold and the

book remained unnoticed for years. Shortly before his death, Gossen withdrew it from circulation and the unsold copies were returned to him. After the author had become famous, Vieweg's successor, Prager, bought this stock from Gossen's nephew, a professor of mathematics by the name of Hermann Kortum, and put it on the market again with a new title page, as a 'second edition', in 1889. There is an Italian translation by Tullio Bagiotti and there is now, since 1983, a careful English translation by Rudolph C. Blitz, nicely divided into chapters. The manuscript of a French translation by Walras was apparently lost.

The first known references to Gossen's book were by Julius Kautz (1858/1860), but they only show that their author did not understand the problems Gossen had solved. Slightly more understanding was shown by F.A. Lange, but again in no more than a footnote. Fortunately, Kautz's reference was seen by Robert Adamson, who was able to get hold of a copy and reported its content to Jevons. In the second edition of *The Theory of Political Economy* (1879) Jevons included a generous acknowledgement of Gossen's priority 'as regards the general principles and method of the theory of Economics', which became the ignition point of Gossen's posthumous fame. Though Gossen's name became famous, his book remains largely unread to this day.

At the level of individual behaviour, Gossen's basic theoretical problem concerns optimization with limited resources (references are to the 1889 edition; they are followed by the corresponding references to the English translation, marked T). Resources are first visualized as time (p. 1f.; T ch. 1). The given lifetime has to be allocated to enjoyable activities in such a way that lifetime enjoyment or, in modern terminology, utility, is maximized.

For a given activity, marginal utility is assumed to be a declining function of the time spent on it. In Gossen's words, 'The magnitude of a given pleasure decreases continuously if we continue to satisfy this pleasure without interruption until satiety is ultimately reached' (p. 4f.; T p. 6). This is the postulate Wilhelm Lexis (1895) christened 'Gossen's First Law'. *In itself,* it was neither new nor profound. W.F. Lloyd had expressed it 20 years earlier just as clearly, it had a long ancestry reaching back to Bentham, the French 'subjectivists', Daniel Bernoulli, and the scholastics, and it is essentially commonplace. To simplify, Gossen assumes marginal utility curves to be linear. It is important to note that Gossen's curves do *not* describe the decline in the marginal utility of a good as its quantity increases, but the decline in the utility from the marginal unit of resources as the quantity of resources is increased. While this facilitated the analysis in some respects, it became a crucial handicap in others. Gossen realized that each of these marginal utility functions must be thought of as being derived by solving a sub-optimization problem, inasmuch as time allocated to a given activity must be spent in the most enjoyable way, probably with interruptions. However, his analysis of this difficult sub-problem, though original and suggestive, remained incomplete and unsatisfactory, leaving much to do for future research on the allocation of time.

Gossen recognized at once that a necessary condition for the optimal allocation of resources is the equality of the marginal utilities in different activities. This is 'Gossen's Second Law', which he had printed in heavy type: 'The magnitude of each single pleasure at the moment when its enjoyment is broken off shall be the same for all pleasures' (p. 12; T p. 14). This theorem is Gossen's principal claim to fame. In it he had no forerunners. It was the key that opened the door to a fruitful analytical use of the First Law and thus initiated the 'marginal revolution' in the theory of value.

The third stage of Gossen's analysis is reached with the introduction of exchange (p. 80f.; T ch. 7). Gossen begins with the bilateral case. He immediately perceives that there are many different opportunities for mutually beneficial exchange, but his discussion of these possibilities is, understandably, inconclusive. As a necessary condition for optimal exchange he postulates that the marginal utilities must be equalized between individuals for each product. While this formulation requires both cardinality and interpersonal comparability of utility, its economic substance, since it can be expressed in terms of 'marginal

rates of substitution, is independent of these assumptions. The concept of a 'contract curve', however, is not used. The statement that each individual would usually be willing to forego a portion of what he receives suggests some notion of consumer's surplus.

The analysis is finally extended to market exchange, where each individual can exchange goods and effort at parametrically given prices, expressed in a common numéraire called money. We thus end up with the optimization problem that became the banner of the 'marginal revolution'. The 'Second Law' can then be expressed by the condition that 'the last atom of money creates the same pleasure in each pleasurable use' (p. 93f.; T p. 109).

The solution to this problem determines the individual's market demand and supply for each product and effort. Gossen also shows how the value of intermediate products can sometimes be derived from that of the final goods, thereby foreshadowing Menger's theory of 'imputation', but he is careful to note that the market mechanism works even where imputation fails (p. 24f.; T p. 28f). If prices are specified at random, aggregate demand and supply will generally differ. Gossen explains how this exerts pressure on prices until all markets are cleared. Prices are thus endogenously determined by general equilibrium. This argument, though concise, is presented in verbal form only. The mathematical formulation of general equilibrium, foreshadowed by Cournot, had to wait for Walras.

In the fourth stage, Gossen introduces rent (p. 102f; T chs. 8–12). If the profundity of an economist can be gauged by his treatment of rent, he comes out near the top. The worker is assumed to own a specific piece of land. Suppose he is now offered the use of land at a superior location, owned by another individual. This does not affect his utility curves, but for the amount of effort for which the marginal utility of effort is just zero he can now earn a higher income. At the same time the marginal disutility curve becomes flatter because the same total enjoyment is now spread over a higher income. In the absence of rent, the superior location would, of course, promise higher income. However, moves to superior locations are not free, but cost rent. This means that at the new location the individual has to earn a certain amount before he can even begin to buy commodities.

What is the maximum rent an individual is willing to pay for a superior location? This 'warranted rent' is reached at the point where total utility at the superior location is equal to the total utility at the original location. Gossen shows algebraically that with rent at the warranted level, superior locations are associated with higher earned income and higher consumption.

While Gossen developed a novel and fruitful way to incorporate rent into a general equilibrium framework, his theory of rent is less rich than von ThÜnen's, published 28 years before. Gossen was no better in reading his predecessors than the later 'marginalists' were in reading Gossen.

The fifth stage introduces capital and interest (p. 114; T ch. 13). The basic question concerns the highest amount of present utility that could be sacrificed for a piece of land with a given annual rent, continuing into the distant future. Gossen finds the answer by discounting the utility of each future rent payment at the appropriate rate of psychological time preference (as we would call it), reflecting uncertainty of expectations (pp. 30, 115; T pp. 35, 134). This promising idea is not successfully exploited, however, and the adaptation of the land paradigm to capital goods remains sketchy. Gossen thinks in terms of land and labour, while capital goods are played down (p. 172; T p. 194). He also makes an effort to determine the optimal amount of saving by the condition that the highest price the individual is willing to pay for a source of rent should be equal to the market price, but he seems to confuse average and marginal concepts and the sense of his argument remains obscure.

In an effort to interpret everyday observations in the light of his theory, Gossen offers an elaborate discussion of the effect of price changes on demand and expenditure. This discussion anticipates a lot of later work on demand elasticities, but it is also cumbersome. The reason is that Gossen's analytical engine, while permitting a brilliantly simple determination of the optimal budget at given prices, is ill suited for the analysis of price change. Since

Gossen's curves, as observed above, relate the marginal utility of expenditure to expenditure, they have to be redrawn after each price change. The insights which Marshall's apparatus made so easy to communicate, remained virtually incommunicable for Gossen. This may be one of the main reasons why his achievement, though at the highest intellectual level, remained sterile. If he had read Cournot, his fate might have been different.

The second part of Gossen's book is largely devoted to social philosophy and policy. It shows its author as a passionate libertarian. Through free markets, mankind would succeed without effort where all socialist planning must fail, namely in reaching the highest possible happiness. Abhorring all forms of protection, Gossen was in favour of free trade, the protection of property rights and a liberal education for both sexes. To prevent fluctuations in the value of money, he advocated a metallic currency and the abolition of paper money. That he also asked for restrictions on child labour and government sponsorship of credit unions seems to indicate that he knew externalities and market imperfections when he saw them. Competitive equilibrium was for him much more than an economic theory or an ideology; it was the gospel, revealing the perfection of a benevolent creator. For him, the 'invisible hand' was not a didactic metaphor, but religion itself. Today, this apotheosis of competition, in language closer to a revival meeting than to scientific discourse, strikes one as bizarre.

Major sources of inefficiency, Gossen thought, were distortions in the allocation of land, preventing land from actually being used by the potentially most efficient user. To correct this defect, he proposed that the government use borrowed money to buy land on the free market and then lease it to the highest bidder (p. 250f.; T ch. 23). Since governments differ from individuals by (1) being immortal, (2) having a higher credit rating, and (3) a lower time preference, such a scheme, he argued, would actually improve government wealth, and the initial debt could eventually be repaid out of rising rent income. For a given year, the scheme would be viable if the price paid by government for a piece of land did not exceed the sum of the rent and the annual increase in the value of land, capitalized at the market rate of interest. However, Gossen was not a 'land socialist'; he was not concerned about 'land monopoly' and the 'socialization of rent'. His objective was the correction of a market imperfection without any limitation of property rights.

Gossen, though perhaps not quite a genius, had a brilliant, original and precise mind. With his one book, he moved constrained optimization into the centre of the theory of value and allocation, where it has since remained. With respect to economic content, his was probably the greatest single contribution to this theory in the 19th century. He failed, however, to develop the basic principle into a usable analytical engine. As a consequence, the so-called 'founders' of the modern theory of value had to rediscover those principles before they could proceed with their engineering work.

Selected Works

1854. *Entwickelung der Gesetze des menschlichen Verkehrs, und der daraus fliessenden Regeln für menschliches Handeln.* Braunschweig: Vieweg. Reprinted, Amsterdam: Liberac, 1967. 2nd ed. Berlin: Prager, 1889. 3rd ed. (with introduction by F.A. Hayek), Berlin: Prager, 1927. Italian translation by T. Bagiotti as *Ermanno Enrico Gossen, Sviluppo delle leggi del commercio umano,* Padua: CEDAM, 1950. English translation by R.C. Blitz (with introductory essay by N. Georgescu-Roegen) as *The laws of human relations and the rules of human action derived therefrom*, Cambridge, MA: MIT Press, 1983.

Bibliography

Bagiotti, T. 1957. Reminiszenzen anlässlich des hundertsten Jahrestages des Erscheinens des Buches von Gossen. *Zeitschrift für Nationalökonomie* 17: 39–54.

Bousquet, G.-H. 1958. Un centenaire: l'oeuvre de H.H. Gossen (1810–1858) et savéritable structure. *Revue d'économie politique* 68: 499–523.

Edgeworth, F.Y. 1896. Gossen, Hermann Heinrich. In *Dictionary of political economy*, ed. R.H. Inglis Palgrave, vol. 2. London: Macmillan.

Georgescu-Roegen, N. 1983. Introduction to H.H. Gossen. *The laws of human relations and the rules of human action derived therefrom*. Trans. R.C. Blitz. Cambridge, MA: MIT Press.
Jevons, W.S. 1879. *The theory of political economy*, 2nd ed. London: Macmillan.
Kautz, J. 1858/1860. *Theorie und Geschichte der National-Oekonomik,* 2 vols. Vienna: Gerold.
Krauss, O. 1910. Gossen, Hermann Heinrich. In *Allgemeine Deutsche Biographie*, vol. 55. Leipzig: Duncker & Humblot.
Lange, F.A. 1875. *Die Arbeiterfrage. Ihre Bedeutung für Gegenwart und Zukunft,* 3rd ed. Winterthur: Bleuler-Hausheer.
Lexis, W. 1895. Art. Grenznutzen. *Handwörterbuch der Staatswissenschaften*, vol. 1 (Supplement). Jena: Fischer.
Liefmann, R. 1910. Hermann Heinrich Gossen und seine Lehre. Zur hundertsten Wiederkehr seines Geburtstages am 7. September 1910. *Jahrbücher für Nationalökonomie und Statistik* 40: 483–498.
Neubauer, J. 1931. Die Gossenschen Gesetze. *Zeitschrift für Nationalökonomie* 2: 733–753.
Pantaleoni, M. 1889. *Principii di economia pura*. Florence: G. Barbéra. *Pure economics* Trans. T.B. Bruce. London: Macmillan, 1898.
Riedle, H. 1953. *Hermann Heinrich Gossen 1810–1858. Ein Wegbereiter der modernen ökonomischen Theorie*. Winterthur: Keller.
Walras, L. 1885. Un économiste inconnu: Hermann-Henri Gossen. *Journal des Economistes.* Reprinted in L. Walrus, *Etudes d'économie sociale.* Lausanne: Rouge, 1896.

Gournay, Jacques Claude Marie Vincent, Marquis de (1712–1759)

Peter Groenewegen

Keywords

Cantillon, R.; Forbonnais, F. V. D.; Free trade; Gournay, Marquis de; Laissez faire; Morellet, Abbe de; Physiocracy; Trudaine, D.; Tucker, J.; Turgot, A. R. J.

JEL Classifications

B31

French economist, merchant and government official, Gournay was born at St. Malo in 1712. After a long career as merchant, spent largely in Cadiz (1729–44), his partner's death in 1746 permitted his retirement two years later from active trade and his entry into public life and more serious research into economics. Gournay has been traditionally associated with the propagation in France of free trade ideas such as deregulation of colonial trade, abolition of the guilds and of the system of government inspection of manufactures, aspects of his work illustrated by the important place generally assigned to him in the history of the phrase, *laissez faire*, *laissez passer* (Schelle 1897, pp. 214–17). Turgot (1759, pp. 30–2) has noted, however, that his free trade position should be qualified and in addition that, unlike the Physiocrats, he accorded an important role in economic development to industry and trade as well as agriculture. He has therefore sometimes been described as the founder of a separate non-Physiocratic free trade school, whose members, among others, included Turgot, Morellet and Trudaine. Apart from *Observations sur l'agriculture, le commerce et les arts de Bretagne* (1757), only his notes accompanying the translation of Child (1754), now edited by Tsuda (1983), appear to have survived. His long friendship with Turgot exerted some influence on the latter's economics, partly because Turgot accompanied Gournay on his tours of inspection of industry between 1753 and 1756. Gournay's most important contribution to French economics seems to have been the encouragement he gave to the study of English economics literature. With Butel-Dumont he had himself translated Child and Culpeper (1754), he encouraged Forbonnais to abridge King's *The British Merchant*, Turgot to translate one of Tucker's pamphlets and, most importantly, may have been responsible for the publication of Cantillon's Essay in 1755 (Morellet 1821, pp. 36–7). His death in 1759 provided the occasion for Turgot's eulogy on which much of the information about his life and work is based, though as Ashley (1900, p. 306) warns, there are reasons for being hesitant in

accepting Turgot's eulogy (1759) 'as evidence of Gournay's opinions'.

Bibliography

Ashley, W.J. 1900. Gournay. In *Surveys, historic and economic*, ed. W.J. Ashley. London: Longman & Co.

Child, J. 1754. *Traités sur le commerce et sur les avantages qui résultent de la reduction de l'intéret de l'argent par Josias Child . . . avec un petit traité contre l'usure par Thomas Culpeper*. Traduit de l'Anglois [by Gournay and Butel-Dumont], Amsterdam/Berlin.

Morellet, l'Abbé. 1821. *Mémoires de l'Abbé Morellet*. Paris: Librairie Française.

Schelle, G. 1897. *Vincent de Gournay*. Paris: Guillaumin.

Tsuda, T. 1983. *Josiah Child, Traités sur le commerce de Josiah Child, avec les remarques inédites de Vincent de Gournay*. Tokyo: Kinokuniya Company.

Turgot, A.R.J. 1759. In praise of Gournay. In *The economics of A.R.J. Turgot,* ed. and Trans. P.D. Groenewegen, The Hague: Nijhoff, 1977.

Government Budget Constraint

Eric M. Leeper and James M. Nason

Abstract

The government budget constraint is an accounting identity linking the monetary authority's choices of money growth or nominal interest rate and the fiscal authority's choices of spending, taxation, and borrowing at a point in time and across time. The intertemporal links create a rich set of possible outcomes from standard macro policy experiments. Taking the government budget constraint seriously can overturn some widely held beliefs about policy effects.

Keywords

Barro, R.; Bond–money ratio; Endowment economy; Euler equation; Fiscal policy; Fiscal theory of the price level; Fisher relation; Government budget constraint; Household budget constraint; Inflation; Markov processes; Monetary policy; Money supply; Ricardian equivalence; Ricardo, D.

JEL Classifications

E6

The government budget constraint is an accounting identity linking the monetary authority's choices of money growth or nominal interest rate and the fiscal authority's choices of spending, taxation, and borrowing at a point in time. Whenever borrowing is the source of some fiscal financing, the government budget constraint also serves to link current monetary and fiscal choices to expected future monetary and fiscal policy variables. This intertemporal dimension creates a rich set of possible impacts of routine macro policy actions, as current or future policies can be expected to adjust to satisfy the government budget, along with other equilibrium conditions. Taking the government budget constraint seriously can overturn some widely held beliefs about policy effects.

The notion that current government policy has intertemporal implications goes back to Barro (1974), who revived ideas associated with Ricardo (1821). Traditional Keynesian models, in contrast, mostly ignored the impact of the government budget constraint on allocations and prices until the work of Christ (1967), 1968 see Sims (1998) for a review and extensions). Hansen et al. (1991) show that identification of the responses of allocations and prices to changes in the government budget constraint require specification of the economic primitives of preferences, technology and market structure.

The modern treatment of the government budget constraint begins with Sargent and Wallace (1981). They show that, when the primary fiscal surplus is fixed, an open-market sale of debt, and contraction of base money, produces higher future inflation. This stunning result arises because, with fiscal policy fixed, faster money supply growth is the only policy expected to balance future government budget constraints. A related but

different mechanism by which the government budget constraint can restrict the equilibrium price level, namely, the 'fiscal theory of the price level', is developed by Leeper (1991), Sims (1994), Woodford (1995, 2001) and Cochrane (1999), among others. That theory demonstrates that, under certain assumptions on policy behaviour, debt-financed cuts in lump-sum taxes can stimulate aggregate demand, in apparent contradiction of Ricardian equivalence.

This article uses endowment and growth economies to study the restrictions the government budget constraint imposes on the intertemporal trade-offs between current and future monetary and fiscal policies. The endowment economy allows us to depict the policy trade-offs associated with a bond-financed tax cut, holding government spending fixed. We show that the effects of policy changes depend on current and expected future monetary and fiscal policies that are consistent with the government budget constraint at each date. Although we illustrate these points for a bond-financed tax cut, analogous results hold for an open-market operation. Implicit in the analyses is that the expected discounted value of real government debt has no value at the infinite horizon; that is, a transversality condition for government debt holds at the infinite horizon. This is a sufficient condition for an equilibrium to exist.

Model Primitives

The models are variations of Sidrauski (1967) and share the following features: perfect foresight, a representative, infinitely lived household with utility defined over consumption, c_t, and real balances, M_t/P_t, $U(c_t, M_t/P_t) = u(c_t) + v(M_t/P_t)$, and nominal one-period government bonds, B, paying net nominal interest of i. The models also have in common two equilibrium conditions that stem from optimal household choices: a portfolio balance expression

$$\frac{v'(M_t/P_t)}{u'(c_t)} = \frac{i_t}{1+i_t}, \tag{1}$$

and optimality of bond choices, a Fisher relation, represented by the Euler equation

$$1 = \beta(1+i_t)\left[\frac{u'(c_{t+1})}{u'(c_t)}\frac{P_t}{P_{t+1}}\right], \tag{2}$$

where $0 < \beta < 1$ is the household's discount factor.

The structure of the government balance sheet, revenue sources, and expenditure process is also common across the models we examine. The government chooses sequences of $\{M_t, B_t, T_t, z_t\}$ to finance purchases of goods and services, g_t, and transfer payments, z_t, to satisfy the government budget constraint

$$g_t + z_t = T_t + \frac{M_t - M_{t-1}}{P_t} + \frac{B_t - (1+i_{t-1})B_{t-1}}{P_t}, \tag{3}$$

where T_t denotes total tax revenues. Government spending is specified as shares of output: $g_t = s_t^g y_t$ and $z_t = s_t^z y_t$. The government budget constraint Eq. (3) has the present value form

$$\frac{B_{t-1}}{P_{t-1}} = E_t \sum_{j=0}^{\infty} \prod_{l=0}^{j} \left(\frac{\pi_{t+l}}{1+i_{t-1+l}}\right)\left[T_{t+j} - g_{t+j} - z_{t+j} + s_{t+j}\right], \tag{4}$$

where $s_t = (M_t - M_{t-1})/P_t$ is seigniorage revenues. To arrive at Eq. (4), the infinite-horizon transversality condition for debt from the household's optimization problem has been imposed: $\lim_{q\to\infty} E_t \beta^q u'(c_{t+q}) \prod_{j=0}^{q} \left(\frac{\pi_{t+l}}{1+i_{t+1+l}}\right) \frac{B_{t+q}}{P_{t+q}} = 0$. This is the relevant sufficient condition because it forces the household to expect that it cannot postpone consumption, hold government bonds for ever, and raise lifetime utility (see Becker and Boyd 1997, for good economic intuition). It is important to note that in stochastic models the transversality condition need not hold always and everywhere along equilibrium paths, as it does in perfect

foresight equilibria. Rather, it holds only in *expectation* (see Kamihigashi 2005, for discussion and examples).

Endowment Economy

It is useful to study an endowment economy because it draws out the role of the government budget constraint in macroeconomic analyses. The household budget constraint is

$$c_t + \frac{M_t + B_t}{P_t} \leq y_t + z_t + \frac{M_{t-1} + (1 + i_{t-1})B_{t-1}}{P_t}, \quad (5)$$

where y is the endowment of goods each period and we have set $T_t = 0$, for all t, so $z^t > 0$ (< 0) represents lump-sum transfers (taxes). Output and government purchases are constant, so $y_t = y$ and $g_t = s^g y$, which implies that in equilibrium consumption is a constant share of GDP, $c_t = c = (1 - s^g)y$. Thus, the equilibrium real interest rate equals a constant, $1/\beta$, the Fisher relation reduces to $1 + i_t = \beta^{-1}\pi_{t+1}$ (where $\pi_{t+1} = P_{t+1}/P_t$), and money demand varies only with the nominal interest rate, $v'(M_t/P_t) = u'(c)[i_t/(1 + i_t)]$.

We focus on circumstances in which the economy is in a stationary equilibrium at dates $s > t$, but starts from a different equilibrium at date t. Denote money growth by $\rho_t = M_t/M_{t-1}$. Assume tax and monetary policies are fixed in the future stationary equilibrium: $s_s^z = s^z$ and $\rho_s = \rho$ for $s > t$; at date t, however, policies may be different: $s_s^z \neq s^z$ and $\rho_s \neq \rho$.

In the stationary equilibrium with constant real money balances, inflation depends only on money growth, $\pi = \rho$, which implies the Fisher relation is $1 + i_s = \beta^{-1}\rho_{s+1}$, $s \geq t$. Stationary real money balances become $M_s/P_s = h(\rho_{s+1})$, for dates $s \geq t$.

We derive two versions of the government budget constraint that describe the trade-offs among current and future monetary and fiscal policies that arise in equilibrium. By imposing equilibrium prices on the government budget constraint Eq. (3), we obtain

$$\frac{h(\rho)}{y}\left[1 - \frac{1}{\rho_t} + \frac{B_t}{M_t} - \frac{1 + i_{t-1}}{\rho_t}\frac{B_{t-1}}{M_{t-1}}\right] = s^g + s_t^z. \quad (6)$$

For given future expected policies, expression Eq. (6) reports the feasible trade-offs among current (date t) policies, when initial liabilities are $(M_{t-1}, (1 + i_{t-1})B_{t-1})$. On the assumption that future policy is anticipated (i.e.,$1 + i = \beta^{-1}\rho$), the government budget constraint is

$$\frac{h(\rho)}{y}\left[1 - \frac{1}{\rho_t} + \left(1 - \frac{1}{\beta}\right)\frac{B}{M}\right] = s^g + s^z, \quad (7)$$

along the equilibrium path for dates $s > t$, given $B_t/M_t = B/M$. Note that the bond–money ratio is constant in the stationary equilibrium. Conditional on the state of government indebtedness, Eq. (7) describes the trade-offs among future policies that are consistent with equilibrium.

Policy Analysis

In the policy experiments we consider, government purchases, s^g, are held fixed. The experiments take the form of an initial cut in taxes at date t (negative s_t^z becomes larger in absolute value), which is financed by sales of nominal bonds. We consider three alternative responses of current and future policies that satisfy Eqs. (6) and (7). The analysis traces the effects of each specification of policy behaviour on the price level and inflation.

Policy 1

For policy experiment 1, suppose current and future money growth, (ρ_t, ρ), are held fixed. This policy, together with the money demand relation, Eq. (1), and Fisher relation, Eq. (2), peg the nominal interest rate at $1 + i = \rho/\beta$ and fix equilibrium real balances at $h(\rho)$. Neither the initial price level, P_t, nor the stationary inflation rate, π, changes. A reduction in taxes today is consistent with equilibrium if nominal debt expands to satisfy the government budget constraint Eq. (6) with fixed money growth. This raises B_t/M_t, which, by

the government budget constraint Eq. (7), forces future taxes to rise sufficiently to service the new, higher level of government indebtedness. This mix of policies yields Ricardian equivalence: the timing of taxes and debt is irrelevant for equilibrium allocations and prices. The policies also imply monetary policy is independent of fiscal considerations, as the quantity theory of money maintains. Of course, as this exercise illustrates, the quantity theory requires specific fiscal behaviour.

Policy 2

In the second experiment, the central bank credibly pegs the nominal interest rate by fixing future money growth, ρ, and the fiscal authority credibly fixes future taxes. Can this be an equilibrium? With future policies fixed, the anticipated budget constraint Eq. (7) implies current policies cannot alter government indebtedness in the future, which is summarized by B/M. Since the expansion in nominal debt cannot be transformed into future higher real debt, P_t must rise in proportion to B_t. However, a pegged nominal interest rate fixes real money balances. The result is that the current money stock must expand in proportion to the increase in prices, which ensures B_t/M_t is unchanged in the date t budget constraint Eq. (6).

The central bank loses control of the current money stock and the price level in this experiment. Changes in these variables are governed by fiscal needs that are beyond the central bank's direct control. A pegged nominal rate subordinates current monetary policy to fiscal needs, but this is not 'monetization of deficits' in the usual sense of printing money to purchase newly issued government debt. Instead, the expansion in money is a passive adjustment of the money supply to clear the money market at the prevailing interest rate and price level. The monetary expansion is given by $dM_t = dB_t/(B_t/M_t)$, making clear that monetary accommodation varies inversely with the level of indebtedness. This exercise corresponds to the fiscal theory of the price level as described by Leeper (1991), Sims (1994), and Woodford (1995). The precise result relies on government debt being sold at par, as Cochrane (2001) observes. If government debt is sold at a discount, bond prices may absorb some of the adjustment to equilibrium, which pushes the price level effects into the future.

Policy 3

The third experiment has the central bank fix current money growth, ρ_t, while the fiscal authority continues to hold future taxes, s^z, constant. It remains feasible for current policy to imply more debt in the future because the anticipated increase in debt service forces future money growth and inflation to rise. The date t response is seen in a higher nominal interest rate and reduced real money balances driven by an increase in P_t to clear the money market, which follows from a fixed M_t. Beyond date t, debt service is financed by higher inflation and seigniorage – 'inflation tax' on nominal assets – revenues. Again, with future net-of-interest fiscal deficits held fixed at $s^g + s^z$, monetary policy is constrained by fiscal needs. In this case, the central bank loses control of future inflation. Sargent and Wallace (1981) employ these assumptions about policy in their classic 'unpleasant monetarist arithmetic' example.

A Growth Economy

A growth model with elastic capital supply, inelastic labour supply, and a distorting income tax extends the analysis by adding interesting intertemporal margins. The model consists of a representative household, a firm that produces the single consumption good, and a government (see, Gordon and Leeper (2006) for related analysis). Assume physical capital depreciates completely after one period. Output is allocated to consumption, capital, k_t, and government purchases of goods, with the technology $f(k_{t-1})$ generating output, y_t, where $f(0) = 0, f'(k_{t-1}) > 0$, and $0 \geq f''(k_{t-1})$. Capital share's of production is denoted by σ. The economy is closed with the aggregate resource constraint

$$c_t + k_t + g_t = f(k_{t-1}). \tag{8}$$

A competitive firm rents capital at rate r from the household and pays taxes levied against

sales of goods, which determine the profit function, $D_t = (1 - \tau_t)y_t - (1 + r_t)k_{t-1}$. Profit maximization yields the after-tax factor price $1 + r_t = (1 - \tau_t)f'(k_{t-1})$.

The household supplies labour inelastically, owns the firm, and receives factor payments. Subject to the budget constraint

$$c_t + k_t + \frac{M_t + B_t}{P_t} \leq (1 + r_t)k_{t-1} + D_t + z_t + \frac{M_{t-1} + (1 + i_{t-1})B_{t-1}}{P_t}, \tag{9}$$

the household maximizes the expected discounted value of its infinite horizon utility function, given P_t , i_t , τ_t, and the initial conditions ($k_{-1} > 0$, $M_{-1} + (1 + i_{-1})B_{-1} > 0$). Government behaviour is unchanged from the endowment economy, but tax revenue is $T_t = \tau_t(1 + r_t)k_{t-1}$.

Equilibrium

We recover an explicit characterization of the model's equilibrium with $u(c_t) = \ln(c_t)$ and $v(M_t/P_t) = \ln(M_t/P_t)$. After imposing transversality conditions for capital, debt and money, equate the supply and demand for capital to find the solution

$$k_t = \left(1 - \frac{1}{\eta_t}\right)(1 - s_t^g)f(k_{t-1}), \tag{10}$$

where $\eta_t \equiv E_t \sum_{i=0}^{\infty} (\sigma\beta)^i \prod_{j=0}^{i-1} \left(\frac{1 - \tau_{t+j+1}}{1 - s_{t+j+1}^g}\right)$

Money market equilibrium sets money supply to money demand to yield

$$\frac{M_t}{P_t} = c_t\mu_t, \tag{11}$$

where $\mu_t \equiv E_t \sum_{i=0}^{\infty} \beta^i \prod_{j=0}^{i-1} \frac{1}{\rho_{t+j+1}}$ Note that μ and η completely summarize what agents need to know to form rational expectations. Since Eqs. (8) and (10) imply a decision rule for consumption, equilibrium real balances can be expressed in terms of their opportunity cost, $1/\mu_t = i_t/(1 + i_t)$, the transactions they help finance, $c_t + k_t$, and expected fiscal policies:

$$\frac{M_t}{P_t} = \frac{1}{\eta_t}\left(\frac{i_t}{1 + i_t}\right)^{-1} c_t + k_t. \tag{12}$$

With $c + k$ serving as a scale variable, expression Eq. (12) is a conventional money demand function except for the dependence on expected fiscal policies. Expectations about future fiscal policies are essential to tie down the equilibrium. This is a key to the dynamics of the growth model and the impacts of fiscal policy on the current equilibrium.

Equilibrium requires that current and future policies satisfy the government's budget constraint and that agents' expectations of policy are consistent with equilibrium. This creates interactions between current and future policies. As before, we distill the analysis down to two periods – now and the future. Fix current and future government spending shares $\{s_t^g, s_t^z\}$, for all t, and assume future money growth and tax rates are constant $\rho_s = \rho$, $\tau_s = \tau$, $s > t$. Current policies, however, may differ: $\rho_t \neq \rho$, $\tau_t \neq \tau$.

The government budget constraint can be expressed entirely in terms of current and expected policies. In period t, the constraint is

$$\left[\frac{\rho_t - 1}{\rho_t} + \frac{B_t}{M_t} - \frac{1 + i_{t-1}}{\rho_t}\frac{B_{t-1}}{M_{t-1}}\right]\frac{\mu_t}{\eta_t} = \frac{s_t^g + s_t^z - \tau_t}{1 - s_t^g}. \tag{13}$$

Given policy expectations are embedded in μ_t/η_t and initial government indebtedness is summarized by $(1 + i_{t-1})B_{t-1}/M_{t-1}$, expression Eq. (13) reports the equilibrium trade-offs among current policies.

Equilibrium trade-offs between current and future policies are given by the state of government indebtedness. We use the budget constraint Eq. (13) to develop this idea for the growth model. Shift the timing of Eq. (13) forward one period and assume future interest liabilities are correctly

G

anticipated at t by substituting the expression for the equilibrium nominal return $1 + i_t$. Given the bond–money ratio is constant at $B/M = B_t/M_t$ in the stationary equilibrium, there can be no net additions to debt in the future. Dropping the time subscript for variables dated $t + 1$ and imposing equilibrium yields

$$\frac{\mu}{\eta}\left[\left(1-\frac{1}{\beta}\right)\frac{B}{M}+\left(\frac{\rho-1}{\rho}\right)\right] = \frac{s^g + s^z - \tau}{1 - s^g}. \tag{14}$$

Equation (14) describes the trade-offs among future policies that are consistent with fixed μ_t/η_t being an equilibrium. The trade-offs represented by Eqs. (13) and (14) tie together current policies and expectations of future policies. Any change in policy at date t that requires a change in μ_t/η_t must be accompanied by a change in policy in the future that is consistent with revised values of μ_t/η_t, conditional on the level of government debt B/M.

Policy Analysis

As for the endowment economy, we study the current and future responses of fiscal and monetary policy to a date t debt finance tax cut in the analysis that follows.

Policy 1
Hold current and future money growth fixed at (ρ_t, ρ). This policy pegs the nominal interest rate by fixing μ_t but it does not fix real money balances unless η_t is also constant. Since new debt issued to finance the tax reduction raises B_t/M_t, a higher level of debt is carried into the future. To clear the government budget constraint in the future, budget constraint Eq. (14) implies future taxes must rise. Higher taxes reduce the return on capital (a lower η) and induce substitution from real to nominal assets, which includes money. Equilibrium in the money market requires the current price level to fall. The source of the non-Keynesian reduction in inflation is the link between current policy (that is, the fiscal expansion) and the expectation that future policy will expand government debt.

Policy 2
Fix both future money growth and future taxes at (ρ, τ). By assumption, all future policies are constant in the face of the current tax cut. Current policies must adjust to ensure the real value of debt in the future is unchanged, as was true when this policy was applied to the endowment economy. The real value of debt remains unchanged because the current money stock rises by the amount that the current budget constraint Eq. (13) dictates is needed to maintain the pre-tax cut level of B_t/M_t. The monetary expansion necessary to maintain equilibrium is sufficient to produce additional future seigniorage (that is, the level of the money supply rises). Since the fixed rate of money growth is just enough to pay for the increased debt service and with equilibrium real money balances fixed by constant future policies, Eq. (12) predicts the current price level rises in proportion to the increase in M_t. Gordon and Leeper (2006) label this 'the canonical fiscal theory exercise'.

The implications of the fiscal theory contrast with the tax cut of policy 1. The bond-financed tax cut is pure fiscal policy in the sense that it is *independent* of the path of the money stock. It also reduces nominal spending and the price level. An essential aspect of the fiscal theory is that the current money stock adjusts passively to clear the money market, raising nominal demand and the price level. If the policy authorities peg the nominal interest rate and fix future taxes without reference to the rest of the economy, higher prices are inevitable consequences of a tax cut. This is an illustration of the fiscal theory.

Policy 3
The fiscal authority holds future taxes constant and the central bank fixes current money growth. If future money growth rises sufficiently to generate the seigniorage revenue to service the new debt, an expansion in current debt can be carried into the future. Expected inflation increases, which lowers the expected return on money (that is, μ falls), decreases money demand, raises the

price level, and contributes to higher future inflation. The change in future money growth depends, of course, on future B/M, which drives the change in debt service.

Concluding Remarks

The equilibria described in this article can easily be couched in terms of arbitrary sequences of policy variables. It has become increasingly popular, however, to endow policy authorities with simple rules that make the policy instrument a time-invariant function of only a few variables that are not directly related to the actions of other policy institutions (that is, the interest rate–monetary policy rules studied in Taylor 1999). Although this approach has the advantages of being interpretable and tractable, it runs the risk of oversimplifying policy behaviour. For example, it is difficult to square simple time-invariant policy functions with the observation that policy regimes can, and do, change, sometimes because of the interactions of different policymakers.

A natural extension of simple rules allows feedback parameters to take on finitely many values ('regimes') whose evolution is governed by a Markov process. Relative to simple rules, this extension produces a far richer set of expectations of future policy variables, a generalization that can overturn some of the principles guiding macro policy research that have been obtained from simple rules (see Davig et al. 2004, and Davig and Leeper 2005).

Markov switching of policy rules has also generalized the test of the long-run sustainability of fiscal policy proposed by Hamilton and Flavin (1986). Davig (2005) finds expansionary and contractionary regimes in US government debt that nonetheless yield a stochastic process for discounted debt with an unconditional expected value equal to zero in the long run.

Theoretical work that takes seriously the restrictions imposed by the government budget constraint has established some important and surprising results. In light of these theoretical findings, it is remarkable how little applied work on monetary and fiscal policy treats the government budget constraint with equal seriousness. This is an open area of research.

See Also

- ▶ Fiscal Theory of the Price Level
- ▶ Hyperinflation
- ▶ Inflation Dynamics
- ▶ Monetary and Fiscal Policy Overview
- ▶ Optimal Fiscal and Monetary Policy (With Commitment)
- ▶ Optimal Fiscal and Monetary Policy (Without Commitment)

G

Bibliography

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82: 1095–1117.

Becker, R., and J. Boyd III. 1997. *Capital theory, equilibrium analysis, and recursive utility*. Malden: Blackwell.

Christ, C. 1967. A short-run aggregate-demand model of the interdependence and effects of monetary and fiscal policies with Keynesian and classical interest elasticities. *American Economic Review* 57: 434–443.

Christ, C. 1968. A simple macroeconomic model with a government budget restraint. *Journal of Political Economy* 76: 53–67.

Cochrane, J. 1999. A frictionless view of U.S. inflation. In *NBER macroeconomics annual 1998*, ed. B. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

Cochrane, J. 2001. Long-term debt and optimal policy in the fiscal theory of the price level. *Econometrica* 69: 69–116.

Davig, T. 2005. Periodically expanding discounted debt: A threat to fiscal policy sustainability? *Journal of Applied Econometrics* 20: 829–840.

Davig, T., and E. Leeper. 2005. Fluctuating macro policies and the fiscal theory. Working paper no. 11212. Cambridge, MA: NBER.

Davig, T., E. Leeper, and H. Chung. 2004. Monetary and fiscal policy switching. Working paper no. 10362. Cambridge, MA: NBER.

Gordon, D., and E. Leeper. 2006. The price level, the quantity theory of money, and the fiscal theory of the price level. *Scottish Journal of Political Economy* 53: 4–27.

Hamilton, J., and M. Flavin. 1986. On the limitations of government borrowing: A framework for empirical testing. *American Economic Review* 76: 808–819.

Hansen, L., W. Roberds, and T. Sargent. 1991. Time series implications of present value budget balance and of martingale models of consumption and taxes. In

Rational expectations econometrics, ed. L. Hansen and T. Sargent. Boulder: Westview Press.
Kamihigashi, T. 2005. Necessity of the transversality condition for stochastic models with bounded or CRRA utility. *Journal of Economic Dynamics and Control* 29: 1313–1329.
Leeper, E. 1991. Equilibria under 'active' and 'passive' monetary and fiscal policies. *Journal of Monetary Economics* 27: 129–147.
Ricardo, D. 1821. *On the principles of political economy and taxation*. 3rd edn. London: John Murray.
Sargent, T., and N. Wallace. 1981. Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis Quarterly Review* 5: 1–17.
Sidrauski, M. 1967. Rational choice and patterns of growth in a monetary economy. *American Economic Review Papers and Proceedings* 57: 534–544.
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.
Sims, C. 1998. Econometric implications of the government budget constraint. *Journal of Econometrics* 83: 9–19.
Taylor, J. 1999. *Monetary policy rules*. Chicago: University of Chicago Press.
Woodford, M. 1995. Price-level determinacy without control of a monetary aggregate. In *Carnegie-Rochester conference series on public policy*, ed. B. McCallum and C. Plosser.
Woodford, M. 2001. Fiscal requirements for price stability. *Journal of Money, Credit, and Banking* 33: 669–728.

Government Budget Restraint

Carl F. Christ

Macroeconomic policy analysis has changed since the government budget restraint (GBR) was incorporated into macroeconomic models. The GBR is the requirement that the total of government expenditure for all purposes – interest, other transfer payments, and goods and services – must equal the total of government financing from all sources – taxes, borrowing from the central bank (i.e. printing money), borrowing from others, and net reduction in reserves of assets such as gold, foreign currency or minerals. For simplicity we ignore changes in reserves.

The GBR implies that the authorities cannot exogenously fix the paths of all the macroeconomic policy variables. They can fix all but one, whereupon the economy and the GBR will endogenously determine the path of the remaining one. For example, if paths of tax rates and spending and borrowing are fixed exogenously, then the change in the monetary base cannot be freely chosen: it must equal spending minus taxes minus borrowing. The authorities can decide which policy variable is to be endogenous and can then fix the paths of the rest. When an exogenous change is made in just one policy variable, the endogenous policy variable must change in response, as must other endogenous variables such as income, prices and interest rates.

The most striking consequence of the GBR is that the economy's path of response to an exogenous change in one policy variable depends on which of the other policy variables is chosen to adjust endogenously, both at impact and during the subsequent adjustment period. Consider a balanced-budget equilibrium that is disturbed by an exogenous step-decrease in tax rates. The economy's response path will be different, depending upon whether the endogenous policy variable is government purchases or transfer payments (either of these will continuously balance the budget) or the monetary base or private holdings of government debt (either of these will at least temporarily involve a budget deficit). Not only does the new equilibrium of the system depend on the choice of the policy variable that is to be endogenous; whether the equilibrium is stable or unstable may also depend on that choice. For example, several authors have concluded that the system is likely to be unstable under a particular form of the monetarist rule; namely, when the monetary base, tax rates and government purchases are fixed exogenously and government debt is endogenous (see below).

Though the facts of government financing and of money creation have long been known, they were not incorporated into mathematical models for some time. Modigliani's (1944) influential model and many of its successors adopted Keynes's (1936) liquidity preference equation and did not include bonds, though non-mathematical accounts often treated liquidity preference as describing the substitution between money and bonds in portfolios. Metzler (1951) broke new

ground by introducing an explicit variable to represent assets (equities) that could be exchanged for money in open-market operations, and Patinkin (1956) explicitly included bonds in his models. Shortly thereafter, the GBR began to appear in macroeconomic analysis.

The flavour of GBR analysis is conveyed by a simple model of a closed economy with no real growth, adapted from Christ (1978, 1979) by assuming that prices adjust instantly so as to maintain real income at full capacity. The model contains the GBR of the consolidated central government sector including the central bank, excluding local governments since they cannot print money.

Symbols are as follows. B = number of privately held perpetual government bonds each of which pays \$I a year nominal interest, g = real government purchases of goods and services, H = monetary base (high powered money), P = price level, π = expected inflation rate, r = nominal interest rate, u = marginal tax rate, V = autonomous nominal taxes, and y = real income (both the actual and the capacity level). V is negative; that is, the tax system is progressive. Transfer payments other than interest are assumed zero, but they could easily be handled by defining u and V to be taxes net of transfers.

Consider the model after it has been reduced to three equations in three endogenous variables: P, r and the endogenous policy variable which may be either B, g, H, u or V. All other variables are exogenous, including y and π. One equation is a financial-assets equilibrium condition, similar to the familiar LM equation except that it allows the demand for the real monetary base to depend on private holdings of government bonds as well as on income and the interest rate. When solved for the latter it becomes:

$$1/r = \lambda(y, H/P, B/P) \tag{1}$$

The second equation is the aggregate demand function:

$$y = \phi(P, \pi; B, g, H, u, V) \tag{2}$$

It was obtained by substituting (1) into a familiar IS equation. The partial derivatives Φ_p, Φ_u, and Φ_V are negative; Φ_g, Φ_H and Φ_π are positive; Φ_B is of uncertain sign, but it must be less than Φ_H/r since an open-market purchase increases aggregate demand. The third equation is the GBR:

$$g + B/P = uy + uB/P + V/P + .H/P + .B/rP \tag{3}$$

where a dot above a symbol denotes its derivative with respect to time. The left side of (3) is real government expenditures for goods and services g and for interest B/P. The right side is real government finance; that is, real taxes at the marginal rate u on real income y and on real interest B/P, plus real autonomous taxes V/P, plus H/P (the real value obtained from the issue of base money) plus B/rP (the real value obtained from the issue of bonds).

In order to permit a balanced equilibrium to exist, consider the case where the growth rates of all the nominal exogenous policy variables are the same. Assume that the exogenous expected inflation rate π is equal to that same growth rate (any plausible expectations-formation process will have this property in equilibrium). Then the system can have a dynamic equilibrium path with steady inflation at the rate π.

For simplicity, consider the case where π and these growth rates are zero. Then, at the static equilibrium, H and B and r drop out of (3), which takes the following static form:

$$P = [(1-u)B - V]/(uy - g) \tag{4}$$

Hence from (2) and (4) one can find the static equilibrium values of P and the endogenous policy variable, and the comparative static effect of any exogenous variable upon P. The effect depends on the choice of which policy variable is to be endogenous. For example, the comparative static effect of the monetary base on the price level, $\partial P/\partial H$, is:

$$-\phi_H(1-u)/P\Delta_B \quad \text{if } B \text{ (bonds) is endogenous} \tag{5}$$

$$-\phi_H/\Delta_g \quad \text{if } g \text{ (government purchases) is endogenous} \tag{6}$$

G

$\phi_H/P\Delta_V$
if V (autonomous nominal taxes) is endogenous (7)

Where Δ with a subscript stands for the determint of the linearized system (2) and (4) when the variable in the subscript of Δ is endogenous. In particular:

$$\Delta_B = [(1-u)B - V]\phi_B/P^2 + (1-u)\phi_p/P \quad (8)$$

$$\Delta_g = [(1-u)B - V]\phi_g/P^2 + \phi_p \quad (9)$$

$$\Delta_H = [(1-u)B - V]\phi_H/P^2 > 0 \quad (10)$$

$$\Delta_V = [(1-u)B - V]\phi_V/P^2 - \phi_p/P > 0 \quad (11)$$

The nature and stability of the dynamic path also depend on which policy variable is endogenous. From the GBR (3) it can be seen that if the monetary base H and private holdings of government bonds B are exogenously held fixed during the adjustment period, the system is not dynamic at all: following any exogenous disturbance, the endogenous response is to balance the budget instantaneously. However, if the endogenous policy variable is either B or H, then the GBR (3) is a dynamic equation, and so the system is dynamic. Then its stability depends on whether B or H is the endogenous variable, as follows.

Suppose H is the endogenous variable. Then the GBR (3) shows that the dynamic path of H is given by:

$$\dot{H} = P \cdot [g - uy - V/P + (1-u)B/P] \quad (12)$$

This is stable iff $\partial\dot{H}/\partial H < 0$, that is,

$$\text{iff} [(1-u)B - V]P^{-1}\phi_H/\phi_P < 0.$$

Since $\Phi_H > 0$ and $\Phi_p < 0$, the system is stable when H is the endogenous variable.

Now suppose B is the endogenous variable. Then the GBR (3) shows that the dynamic path of B is given by:

$$\dot{B} = rP \cdot [g - uy - V/P + (1-u)B/P \quad (13)$$

This is stable iff $\partial\dot{B}/\partial B < 0$; that is [using (8)], iff $rP\ \Delta_B/\Phi_p < 0$.

Since the sign of Φ_B is uncertain, the sign of Δ_B in (8) seems uncertain. However, note from (5) that if the effect of the monetary base upon the equilibrium price level is positive when B is endogenous, as is plausible, then the positive sign of Φ_H implies that $\Delta_B < 0$, and hence the system is unstable when B is the endogenous variable. Similar results have been obtained by several others; for example, Tobin and Buiter (1976).

There is a large literature on the GBR, in which the foregoing analysis has been extended in several directions. Steady-inflation equilibrium paths have been considered. So have steady-real-growth equilibrium paths. (In either case, the equilibrium path has a constant real budget deficit. A steady-inflation equilibrium is possible only if the real deficit is not too large to be financed by the inflation tax.) Deviations of output from the capacity level (as in business cycles) have been introduced. The foreign sector has been included. Further, two or more policy variables can be endogenous at the same time, if there is a policy rule governing their joint responses.

Does the Ricardian equivalence theorem of Barro (1974) obviate the need for the GBR? The theorem says that under suitable conditions (including the dubious assumption that the interest on government debt will certainly be covered by future taxes, without inflation or default) bond finance is equivalent to tax finance. It implies that any change in the timing of tax payments will have no effect on private behaviour as long as the present value of tax payments is not altered. As Barro recognizes, this is not true for persons who are at a corner solution, consuming less than they would if they could shift some purchasing power from the future to the present.

But suppose there are never any such persons. The Equivalence Theorem implies that government bonds are not net wealth. According to Sargent (1979, ch. 4) and McCallum (1978), this means that models like (1)–(2) above should be

modified in such a way that, once the time-paths of the monetary base and government purchases are fixed exogenously, the GBR and the proportion of bond finance to tax finance are irrelevant for private behaviour. If this were correct, a government could set its taxes permanently at zero and finance its expenditures, including interest, solely by issuing new debt forever, without affecting the interest rate, prices or output.

Suppose that the interest rate exceeds the economy's growth rate. This is one of the premises of the Equivalence Theorem. It is plausible, since the golden rule of economic growth in Phelps (1965) shows that the economy cannot be on an optimum steady state path if the reverse inequality holds. Then, as Barro (1974, 1976), Sargent and Wallace (1981), and McCallum (1984) recognize, it is not possible to pursue forever a policy of continually borrowing to pay the debt interest. It would eventually make the debt interest exceed the revenue capacity of the tax system. Hence debt interest could not be covered by taxes, and the Equivalence Theorem would fail. The result would be inflation or outright default. Thus the omission of the GBR and government debt can lead to incorrect conclusions.

See Also

- ▶ Public Debt
- ▶ Ricardian Equivalence Theorem

Bibliography

Barro, R. 1974. Are government bonds net wealth? *Journal of Political Economy* 82: 1095–1117.

Barro, R. 1976. Reply to Feldstein and Buchanan. *Journal of Political Economy* 84: 343–349.

Christ, C.F. 1978. Some dynamic theory of macroeconomic policy effects on income and prices under the government budget restraint. *Journal of Monetary Economics* 4: 45–70.

Christ, C.F. 1979. On fiscal and monetary policies and the government budget restraint. *American Economic Review* 69: 526–538.

Keynes, J.M. 1936. *The general theory of employment interest and money*. New York: Harcourt Brace.

McCallum, B.T. 1978. On macroeconomic instability from a monetarist policy rule. *Economics Letters* 1: 121–124.

McCallum, B.T. 1984. Are bond-financed deficits inflationary? A Ricardian analysis. *Journal of Political Economy* 92: 123–135.

Metzler, L.A. 1951. Wealth, saving and the rate of interest. *Journal of Political Economy* 59: 93–116.

Modigliani, F. 1944. Liquidity preference and the theory of interest and money. *Econometrica* 12: 45–88.

Patinkin, D. 1956. *Money, interest, and prices*. Evanston: Row Peterson.

Phelps, E.S. 1965. Second essay on the golden rule of accumulation. *American Economic Review* 55: 793–814.

Sargent, T. 1979. *Macroeconomic theory*. New York: Academic.

Sargent, T., and N. Wallace. 1981. Some unpleasant monetarist arithmetic. *Federal Reserve Bank of Minneapolis Quarterly Review*, Fall, 1–17.

Tobin, J., and W. Buiter. 1976. Long-run effects of fiscal and monetary policy on aggregate demand. In *Monetarism*, ed. Jerome L. Stein. Amsterdam: North-Holland.

Graham, Frank Dunstone (1890–1949)

Randall Hinshaw

Keywords

Ceiling velocity; Commodity-reserve monetary standard; Flexible exchange rates; German hyperinflation; Graham, F. D.; International trade theory; Mill, J. S.; Protection; Tariffs; Taussig, F. W.; Terms of trade; Velocity of circulation

JEL Classifications

B31

Graham was born in Halifax, Nova Scotia, and died in Princeton, New Jersey. He is known mainly for his work in the theory of international trade, and especially for his attack on classical and neoclassical trade theory. He received his doctorate from Harvard, where he came under the influence of Taussig. After teaching at Rutgers and

Dartmouth, he joined the Princeton faculty in 1921, becoming a full professor in 1930. In addition to undergraduate teaching, he taught the Princeton graduate courses in international trade and in monetary theory.

In a path-breaking article (1923b), Graham argued that J.S. Mill, by using a two-country, two-commodity model, had reached erroneous conclusions concerning the effect of changes in international demand on the commodity terms of trade. Mill had reasoned that, within the limits set by comparative cost (limits which he – but not Graham – regarded as improbable cases), an increase in a country's demand for imports would worsen the country's terms of trade. Retaining Mill's assumptions of free trade, costless transportation and constant cost per unit of output, Graham concluded that, when a given commodity is produced by more than one country, the cost structures of the affected countries are locked together and that, therefore, changes in international demand do not affect the equilibrium terms of trade so long as the same commodities continue to be produced by the same countries; instead, within possibly wide limits, international adjustment takes place through shifts in output, the limits occurring when commodities disappear from, or are added to, national production schedules.

To illustrate these points, Graham devised a multi-country, multi-commodity model, with all variables expressed in real terms. Operating with assumed national opportunity-cost ratios, national productive capacities and national demand functions, he was able to derive, by a trial-and-error process using simple arithmetic, an equilibrium solution specifying the commodity terms of trade and each country's consumption, production (if any) and exports or imports of each commodity. In his final work, *The Theory of International Values* (1949), he developed these ideas at length, using illustrations with as many as ten countries and ten commodities. Because of the assumption of costless transportation, domestic (non-traded) goods do not appear in the trade model, but Graham examined their role in international adjustment in his earliest article (1922) and in his 1949 treatise.

Although the Graham model, which assumes full employment, can be used to demonstrate that national and world real output are maximized under free trade, Graham was not a doctrinaire free trader. In an early article (1923a), he made a case for permanent protection for decreasing-cost industries. The article was attacked on various grounds by Knight (1924) and others, but Graham retained the argument in his book, *Protective Tariffs* (1934), which, while critical of most arguments for tariffs, included a chapter on 'Rational Protection'.

In the field of money, Graham's major work was his treatise (1930) on the German hyperinflation after the First World War. Perhaps his most significant conclusion was the concept of 'ceiling velocity'. He found that, in the German case, monetary velocity reached an upper limit which was about 25 times the pre-war normal; thereafter, the German price level rose at approximately the same rate as the German money supply.

Graham had a passionate interest in economic policy. He was an early advocate of flexible exchange rates (on a managed basis), and during the Great Depression he devised various plans to promote recovery. Later, he advocated a commodity-reserve monetary standard as a means of achieving price-level stability and full employment.

An iconoclast with a caustic wit, Graham was an unusually stimulating teacher and had a profound influence on his students, two of whom – T.M. Whitin and L.W. McKenzie – extended his work on the trade model. In a 1953 article which illustrated the model geometrically, Whitin concluded that Graham's work 'anticipated linear programming models by many years', and McKenzie, in a powerful 1954 article employing a theorem from topology, demonstrated what Graham firmly believed but was never able to prove: that his trade model yields an equilibrium for any continuous demand functions and that this solution is unique for the demand functions which Graham actually used.

Selected Works

1922. International trade under depreciated paper: The United States, 1862–79. *Quarterly Journal of Economics* 36: 220–273.

1923a. Some aspects of protection further considered. *Quarterly Journal of Economics* 37: 199–227.
1923b. The theory of international values re-examined. *Quarterly Journal of Economics* 38: 54–86.
1930. *Exchange, prices, and production in hyper-inflation: Germany, 1920–1923*. Princeton: Princeton University Press.
1932a. The theory of international values. *Quarterly Journal of Economics* 46: 581–616.
1932b. *The abolition of unemployment.* Princeton: Princeton University Press.
1934. *Protective tariffs.* New York: Harper & Bros. Reprint edn. Princeton: Princeton University Press, 1942.
1942. *Social goals and economic institutions.* Princeton: Princeton University Press.
1949. *The theoryof international values.* Princeton: Princeton University Press.

Bibliography

Knight, F.H. 1924. Some fallacies in the interpretation of social cost. *Quarterly Journal of Economics* 38: 582–606.
McKenzie, L.W. 1954. On equilibrium in Graham's model of world trade and other competitive systems. *Econometrica* 22: 147–161.
Metzler, L.A. 1950. Graham's theory of international values. *American Economic Review* 40: 301–322.
Whitin, T.M. 1953. Classical theory, Graham's theory, and linear programming in international trade. *Quarterly Journal of Economics* 67: 520–544.
Whittlesey, C.R. 1952. Frank Dunstone Graham, 1890–1949. *Economic Journal* 62: 440–445.

Gramsci, Antonio (1891–1937)

Massimo L. Salvadori

Italian communist and Marxist theorist, Gramsci was born in 1891 in Ales (Sardinia), and died in Rome in 1937.

Gramsci's work acquired national importance in Italy when in 1919, together with A. Tasca, U. Terracini and P. Togliatti, he founded the weekly magazine *Ordine nuovo.* The aim of this publication, under the influence of the Russian revolution, was to disseminate the idea of a proletarian dictatorship based on workers' 'councils' and on the alliance between the workers of northern Italy and the poor peasants of the south. Gramsci was elected member of Parliament (1924–6) and became secretary of the Italian Communist Party (PCI) in 1924. In spite of the fact that he had opposed the left in the CPSU in 1926, Gramsci, in bitter opposition to Togliatti, warned of the danger that Bukharin and Stalin's aim was to crush their opponents and subject the International to Russian national interests. Gramsci was arrested in November 1926 and condemned by the special fascist Tribunal to twenty years' imprisonment. During his time in prison he made notes and kept records which were collected in *Quaderni del carcere* (first published between 1948 and 1951). In 1930 he rejected the theory of 'social fascism' supported by the third International. He later became seriously ill and died in a clinic in Rome in 1937.

While in jail Gramsci was aided in many ways by his friend the eminent economist Piero Sraffa, who had moved to Cambridge, England, in 1927.

Gramsci's Marxist beliefs developed into an anti-positivistic attitude. He was affected by the idealism of Gentile and Croce, and also by the thought of Sorel and Bergson. It seemed to him that Lenin and the Bolshevik Party were the ideological incarnation of the new Marxism, organized into an active political force, in direct contrast to the old deterministic Marxism of the social democratic parties of the Second International.

In 1919–20 Gramsci had believed in the supremacy of the revolutionary intitiative of the 'Workers' Councils'. After 1921, as a result of a deeper understanding of Leninism, he changed his perspective and underlined instead the primacy of the party as interpreter of the revolutionary process.

During the time that he was in prison, he reflected on the causes of the defeat of the Revolution in the West. He wrote in the *Quaderni* that the social, political and cultural differences between the East and West were such that the Russian Revolution could not be adopted as a model to be

copied automatically. In the West the accession to power would have to be preceded by a period of intense political struggle ('war of position') during which the Communist Party (the 'Modern Prince') and the proletariat would have to form a broad front of social alliances and win a wide political and cultural 'consensus' (the theory of 'hegemony').

Gramsci believed that Italy had missed out on the opportunity of producing a national bourgeoisie capable of ensuring the development of a modern society. Italy's inability to solve the problems of the South ('the southern question') bore witness to this. Gramsci believed that it was up to the PCI to change Italian society and, by creating a new socialist order, to accomplish the difficult task of 'national' unification.

Gramsci's beliefs exerted a wide influence on the left, first in Italy, and then in Western Europe. The PCI, which had at the beginning judged him to be a great 'orthodox Leninist', later used Gramsci's 'theory of hegemony' as its main theoretical inspiration for 'Eurocommunism', thus forming a political strategy aimed at surmounting the limits of Leninism.

Gramsci never paid any systematic attention to economic theory. Nevertheless he wrote on it, especially in the *Quaderni* which includes many methodological notes. He was against using the concept of 'laws' according a deterministic pattern both in economics and sociology. In his opinion, only Marxism was able to establish a 'critical' conception of economics. The 'value' – he stated – is the very core of Marxist economic theory, as far as it explains the 'relationship between the worker and the industrial forces of production'. And whereas the bourgeois idea of 'market' is an 'abstract' one, the Marxist idea is related to 'historicism', that is it is based on the consciousness of the social and historical conditions of the market itself, which have to be changed in consequence of the revolutionary process.

Selected Works

1947. *Lettere dal carcere*. Turin: Einaudi. Trans. by Lynne Lawner as *Letters from prison*. New York: Harper & Row, 1973.

1948–51. *Quaderni del carcere*, 6 vols. Turin: Einaudi. New ed, 4 vols. Turin: Einaudi, 1975. Selections Trans. by Quintin Hoare and Geoffrey Nowell Smith as *Selections from the prison note-books*. London: Lawrence & Wishart, 1971.

1954. *L'ordine nuovo*. Turin: Einaudi. Another ed, 1975.

1971. *La costruzione del Partito Comunista*. Turin: Einaudi.

References

Adamson, W.L. 1980. *Hegemony and revolution. A study of Antonio Gramsci's political and cultural theory.* Berkeley: University of California Press.

Buci-Glucksmann, C. 1975. *Gramsci et l'état*. Paris: Fayard.

Clark, M. 1977. *Antonio Gramsci and the revolution that failed*. New Haven: Yale University Press.

Fiori, G. 1966. *Vita di Antonio Gramsci*. Bari: Laterza.

Romeo, R. Risorgimento e capitalismo. Bari: Laterza.

Salvadori, M.L. 1970. *Gramsci e il problema storico della democrazia*. Turin: Einaudi.

Spriano, P. 1967. *Storia del Partito Comunista italiano*, vol. 1. Turin: Einaudi.

Spriano, P. 1977. *Gramsci in carcere e il Partito.* Rome: Editori Riuniti. Trans. by John Fraser as *Antonio Gramsci and the party: The prison years*. London: Lawrence & Wishart, 1979.

Togliatti, P. 1967. *Gramsci*. Rome: Editori Riuniti.

Granger, Clive W. J. (1934–2009)

Timo Teräsvirta

Keywords

Cointegration; Conditional variance models; Degree of integration; Econometrics; Forecasting; Fractional integration; GARCH models; Granger representation theorem; Granger, C.W. J.; Granger-causality; Haavelmo, T.; Long memory models; Maximum likelihood; Nonlinear time series; Non-stationary variables; Reduced rank regression; Spectral analysis; Spurious regressions; Time series analysis; Vector autoregressions

JEL Classifications
B31

Granger, Clive William John, born 4 September 1934, Swansea, Wales. British citizen, knighted in the 2005 New Year's Honours. Emeritus Professor of Economics, University of California, San Diego. Degrees: BA, University of Nottingham 1955, Ph.D., University of Nottingham 1959. Career: Lecturer, then Professor of Applied Statistics and Econometrics, University of Nottingham, 1956–73; Professor of Economics, University of California, San Diego, 1974–2003. Honours and awards: Fellow, Econometric Society 1972; Fellow, American Academy of Arts and Sciences 1994; Fellow, International Institute of Forecasters 1996; Foreign member, Finnish Society of Sciences and Letters 1997; Corresponding Fellow, British Academy 2002; Distinguished Fellow of American Economic Association 2002; Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel 2003 'for methods of analyzing economic time series with common trends (cointegration)'. Honorary degrees: University of Nottingham, Universidad Carlos III de Madrid, Stockholm School of Economics, University of Loughborough, Aarhus University, Aristotle University.

Clive Granger is one of the best-known time-series econometricians of our time. He has contributed to many areas in econometrics. They include the analysis of nonstationary time series, causal relations between economic variables, long memory, nonlinearity, forecasting economic time series, modelling stock prices and volatility, and price formation.

The most important concept that Granger has introduced to econometrics is *cointegration.* It can be viewed as an extension to non-stationary time series of Nobel Laureate Trygve Haavelmo's formulation of an economy as a system of simultaneous stochastic relationships that laid the foundations to modern time-series econometrics. The origins of the concept may be traced to a paper in which Granger and his associate Paul Newbold consider regressing a *random walk* series on another independent random walk series (Granger and Newbold 1974). They pointed out that the classical t-test in such a regression may often suggest a statistically significant relationship between variables although none exists. This result suggested that many relationships found between non-stationary economic variables in static econometric models of the time could in fact have been spurious. It called for a more careful analysis of non-stationary economic time series, and Granger and Newbold also strongly emphasized the importance of dynamic models instead of static ones that did not take the dynamic properties of economic variables into account.

A solution to this spurious regression problem was to model the relationships between economic variables in first differences. This, however, created a problem because these relationships were generally expressed in levels, not differences. Granger's solution to this problem (Granger 1981) may be illustrated by the following regression equation:

$$y_t = \alpha + \beta x_t + \varepsilon_t, \tag{1}$$

where y_t is the dependent variable, x_t the single exogenous regressor, and $\{\varepsilon_t\}$ a white-noise, mean-zero sequence. Granger used the concept of *degree of integration* of a variable. If variable z_t can be made approximately stationary by differencing it d times, it is called integrated of order d, or I(d). Weakly stationary random variables are thus I(0). Many macroeconomic variables can be regarded as I(1) variables: if $z_t \sim$ I(1), then $\Delta z_t \sim$ I(0). Assume now that both $x_t \sim$ I(1) and $y_t \sim$ I(1) in Eq. (1). Then generally the linear combination $y_t - \beta x_t \sim$ I(1) as well. There is, however, one important exception. Many macroeconomic variables can be regarded as I(1) variables: if $z_t \sim$ I(1), then $\Delta z_t \sim$ I(0). It has to do with the fact that for an equation such as (1) to be meaningful it has to be *balanced,* a concept Granger employed in his work. An equation is balanced when its right-hand and left-hand sides are of the same order of integration. Rewrite (1) as

$$y_t - \beta x_t = \alpha + \varepsilon_t.$$

If $\varepsilon_t \sim$ I(0), then $y_t - \beta\, x_t \sim$ I(0), that is, the linear combination $y_t - \beta\, x_t$ has the same statistical properties as an I(0) variable. There exists only one such combination so that coefficient β is unique. In this special case, variables x_t and y_t are called *cointegrated*. This notion generalizes to more than two variables.

The importance of cointegration in the modelling of non-stationary economic series becomes clear in the Granger representation theorem that was first formulated in Granger and Weiss (1983). Consider the following bivariate vector autoregressive (VAR) model of order p:

$$x_t = \sum_{j=1}^{p} \gamma_{1j} x_{t-j} + \sum_{j=1}^{p} \delta_{1j} y_{t-j} + \varepsilon_{1t}$$
$$y_t = \sum_{j=1}^{p} \gamma_{2j} x_{t-j} + \sum_{j=1}^{p} \delta_{2j} y_{t-j} + \varepsilon_{2t},$$

where x_t and y_t are I(1) and cointegrated, and ε_{1t} and ε_{2t} are white noise. The Granger representation theorem states that in this case the system can be written as:

$$\begin{aligned} \Delta x_t &= \alpha_1 (y_{t-1} - \beta x_{t-1}) \\ &+ \sum_{j=1}^{p-1} \gamma_{1j}^{*} \Delta x_{t-j} + \sum_{j=1}^{p-1} \delta_{1j}^{*} \Delta y_{t-j} + \varepsilon_{1t} \Delta y_t \\ &= \alpha_2 (y_{t-1} - \beta x_{t-1}) \\ &+ \sum_{j=1}^{p-1} \gamma_{2j}^{*} \Delta x_{t-j} + \sum_{j=1}^{p-1} \delta_{2j}^{*} \Delta y_{t-j} + \varepsilon_{2t}, \end{aligned} \tag{2}$$

where at least one of parameters α_1 and α_2 deviates from zero. Both equations of the system are balanced since $y_{t-1} - \beta x_{t-1} \sim$ I(0). System (2) is now in *error-correction form* where $y_t - \beta x_t = 0$ defines a dynamic equilibrium relationship between the two economic variables, y and x. While the system consists of two equations, it only has a single equilibrium relationship. More generally, if the system has n variables, the number of these *cointegrating relationships* is less than n. System (2) in disequilibrium but has a built-in tendency to adjust itself towards the moving equilibrium. The coefficients α_1 and β_2 represent the relative strength of this adjustment at any given time. In applications, the equilibrium or long-run relationship represents an economic theory proposition, whereas the remaining variables and parameters describe the short-term dynamic behaviour of the system.

It may be mentioned that linear combinations of non-stationary variables had appeared in dynamic econometric equations prior to Granger's work. Phillips (1957), who coined the term 'error correction', and Sargan (1964) had employed them, but they did not, however, consider the statistical implications of introducing such components into their models.

A first test for cointegration appeared in Granger and Weiss (1983). The idea is best seen from Eq. (1). Variables x_t and y_t are cointegrated if "$\varepsilon_t \sim$ I(0). Estimate parameter β by ordinary least squares and apply a standard unit root test to the residuals. A rejection of the null hypothesis suggests cointegration.

A method for estimating parameters of systems with cointegrated variables was still needed to make the concept applicable. Granger, working jointly with Robert Engle, developed a two-stage estimation method for VAR models with cointegration (Engle and Granger 1987). Consider the following n-dimensional VAR model of order p:

$$\begin{aligned} \Delta \mathbf{x}_t &= \alpha \beta' \mathbf{x}_{t-1} \\ &+ \sum_{j=1}^{p-1} \Gamma_j \Delta \mathbf{x}_{t-j} + \varepsilon_t (t = 1, \ldots, T) \end{aligned} \tag{3}$$

where $\mathbf{x}_t$ is an $n \times 1$ vector of I(1) variables, $\alpha\beta'$ is an $n \times n$ matrix such that the $n \times r$ matrices α and β have rank r, Γ_j, $j = 1,\ldots,p-1$, are $n \times n$ parameter matrices, and ε_t is an $n \times 1$ vector of white noise with a positive definite covariance matrix. If $0 < r < n$, the variables in $\mathbf{x}_t$ are cointegrated with r cointegrating relationships $\beta'\mathbf{x}_t$. If the variables in $\mathbf{x}_t$ are cointegrated, the parameters of (3) can be estimated in two stages. First estimate β or, more precisely, the cointegrating space (β up to a multiplicative constant) using a form of least squares. Then, holding that estimate fixed, estimate the remaining parameters by maximum likelihood. The estimators of a and Γ_j, $j = 1,\ldots,p-1$, are consistent and asymptotically normal. This solution is based on

the fact that the least squares estimator of β is superconsistent: its rate of convergence is faster than that of the estimators of the other parameters. Engle and Granger (1987) also contains a rigorous proof of the Granger representation theorem.

This paper by Engle and Granger is one of the most cited papers in time series econometrics (200 citations annually since its appearance), and it was followed by a flood of applications. Cointegration strongly contributed to the popularity of VAR models suggested by Sims (1980) as a convenient tool for modelling economic variables without strong assumptions originating from economic theory. The ultimate refinement of the statistical theory of cointegrated variables was provided by Søren Johansen (see Johansen 1995, for a summary) who derived the maximum likelihood estimator of β or, more precisely, the space spanned by the r cointegrating vectors in (3), using *reduced rank regression.* He also considered tests for determining the cointegration rank r. It may be mentioned that Granger originally (Granger 1981) defined cointegration using tools from *spectral analysis.* In fact, in 1964 he wrote an early book on the topic jointly with Michio Hatanaka. The book that became a Citation Classic was based on the work carried out when Granger was visiting Princeton University on a Harkness fellowship in the early 1960s.

Granger also undertook very influential research concerning causal relationships between variables. He moved away from deterministic causality ('if A then B') to stochastic one in which time plays a decisive role. Granger's causality definition (Granger 1969a) is based on the prediction accuracy of a stationary variable y. If y can be predicted more accurately with a set of other variables than without them, then these variables are said to cause. This is an operational definition that makes it possible to test the null hypothesis of no causality between economic variables with statistical methods. This may be done in either a single-equation or a system framework. For this reason this concept of causality, nowadays called Granger-causality, has become very popular in applied work. Most of the available tests, however, test for in-sample predictability, whereas Granger has always emphasized the fact that his definition pertains to out-of-sample forecasting and the (non) existence of causality should be tested accordingly. Granger-causality has become an important tool in economic research and policymaking and is also being used in other areas than economics. For more information, see Hendry and Mizon (1999). Granger (1986) established a link between causality and cointegration. If y and x are cointegrated then there is Granger-causality at least in one direction between these variables.

Granger has also been instrumental in starting the econometric research on processes with *long memory.* They have the property that their autocorrelations as a function of the lag length decay at a slower rate than autocorrelations of a linear autoregressive–moving average process in which the decay rate is exponential. Granger and Joyeux (1980) defined a new concept, *fractional integration,* and showed how fractionally integrated processes, stationary or non-stationary, can have long memory. A stochastic variable y_t is fractionally integrated series of order d, I(d), where d need not be an integer, if x_t in

$$x_t = (1 - L)^d y_t$$

where L is the lag operator: $Lyt = y_{t-1}$, is an I (0) variable. Choosing $d = 1$ yields the standard case where $y_t \sim$ I(1). Time series models based on fractional integration have since become popular in econometrics, and long memory in volatility has received plenty of attention in financial econometrics.

Granger is one of the first econometricians interested in *nonlinear time series* and wrote (1978, with Allan Andersen) a book on bilinear time series models. The bilinear model has not found much application in economics, but the book has stimulated more research in this area. Granger has since expanded his interests in nonlinear models, among other things, by generalizing cointegration, originally a linear concept, into nonlinear cointegration. He has also written a book on nonlinear econometric modelling (1993, with Timo Teräsvirta).

Economic forecasting has been one of Granger's main interests throughout his career.

He observed (Bates and Granger 1969b) that combining forecasts from different models often improves the forecast accuracy compared with forecasts from individual models, and proposed forecast weighting schemes for this purpose. This work has prompted a large literature that is still growing. He wrote a book (1970, with Oskar Morgenstern) on forecasting stock markets before that became a topic of broad interest, and a classic textbook with Paul Newbold on economic forecasting (Granger and Newbold 1977). Another forecasting topic that has attracted Granger's interest is the evaluation of forecasts and the role of the forecaster's cost function in both parameter estimation and model evaluation, and he has made important contributions in this area.

One may argue that Granger's research is principally focused on conditional mean models, but he has also contributed to the analysis of conditional variance models. He has extended the standard model of *generalized autoregressive heteroskedasticity* (GARCH) model into the power GARCH model (Ding, Granger and Engle 1993), intended to improve the modelling of volatility in financial time series such as return series of sufficiently high (intra-daily, daily, weekly) frequency. He has also indicated the potential for statistical modelling of the decomposition of stock return series into a sign process with little or no autocorrelation and an absolute return one with strong dependence structure. He has considered forecasting volatility of financial return series, a topic that is of great importance to investors who consider volatility to be a measure of risk (Poon and Granger 2003).

A representative collection of Granger's scientific papers can be found in Ghysels et al. (2001).

See Also

- ▶ Cointegration
- ▶ Forecasting
- ▶ Granger–Sims Causality
- ▶ Spurious Regressions
- ▶ Vector Autoregressions

Selected Works

1964. (With M. Hatanaka.) *Spectral analysis of economic time series.* Princeton: Princeton University Press. Also in French.

1969a. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37: 423–38.

1969b. (With J. M. Bates.) The combination of forecasts. *Operations Research Quarterly* 20:, 451–68.

1970. (With O. Morgenstern.) *Predictability of stock market prices.* Lexington. MA: D.C. Heath.

1974. (With P. Newbold.) Spurious regressions in econometrics. *Journal of Econometrics* 2: 111–20.

1977. (With P. Newbold.) *Forecasting Economic Time Series.* New York: Academic Press. 2nd edn, 1986.

1978. (With A.P. Andersen.) *Introduction to bilinear time series models.* Göttingen: Vandenhoeck & Ruprecht.

1980. (With R. Joyeux.) An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis* 1: 15–30.

1981. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16: 121–30.

1983. (With A.A. Weiss.) Time series analysis of error-correction models. In *Studies in econometrics, time series and multivariate statistics in honor of T.W. Anderson*, ed. S. Karlin, T. Amemiya and L.A. Goodman. San Diego: Academic Press.

1986. Developments in the study of cointegrated variables. *Oxford Bulletin of Economics and Statistics* 48: 213–28.

1987. (With R. F. Engle.) Cointegration and error correction: Representation, estimation and testing. *Econometrica* 55: 251–76.

1993. (With T. Teräsvirta.) *Modelling nonlinear economic relationships.* Oxford: Oxford University Press.

1993. (With Z. Ding and R.F. Engle.) A long memory property of stock market returns and

a new model. *Journal of Empirical Finance* 1: 83–106.
2003. (With S.-H. Poon.) Forecasting volatility in financial markets: a review. *Journal of Economic Literature* 41: 478–539.

Bibliography

Ghysels, E., N.R. Swanson, and M.W. Watson, ed. 2001. *Essays in econometrics. Collected papers of Clive W.J. Granger*. Vol. 1 and 2. Cambridge: Cambridge University Press.
Hendry, D.F. 2004. The Nobel memorial prize for Clive W.J. Granger. *Scandinavian Journal of Economics* 106: 187–213.
Hendry, D.F., and G.E. Mizon. 1999. The pervasiveness of Granger causality in econometrics. In *Cointegration, causality and forecasting. Festschrift in honour of Clive W.J. Granger*, ed. R.F. Engle and H. White. Oxford: Oxford University Press.
Johansen, S. *1995*. *Likelihood-based inference in cointegrated vector autoregressive models*. *Oxford*: *Oxford University Press*.
Phillips, A.W. 1957. Stabilization policy and the time forms of lagged responses. *Economic Journal* 67: 265–277.
Phillips, P.C.B. 1997. The ET interview: Professor Clive Granger. *Econometric Theory* 13: 253–303.
Sargan, J.D. 1964. Wages and prices in the United Kingdom a study in econometric methodology. In *Econometric analysis for national economic planning*, ed. P.E. Hart, G. Mills, and J.K. Whitaker. London: Butterworths.
Sims, C.A. 1980. Macroeconomics and reality. *Econometrica* 48: 1–48.
Teräsvirta, T. 1995. Professor Clive W.J. Granger: An interview for the *International Journal of Forecasting*. *International Journal of Forecasting* 11: 585–590.

Granger–Sims Causality

G. M. Kuersteiner

Abstract

The concept of Granger–Sims causality is discussed in its historical context. There follows a review of the subsequent literature that explored conditions under which the definitions of Granger and Sims are equivalent. The relationship to the potential outcomes framework is explored in light of recent developments in the literature.

Keywords

Block recursive structure; Causality in economics and econometrics; Conditional independence; Conditional probability; Covariance stationary processes; Equivalence relationships; Granger non-causality; Granger, C.; Granger–Sims causality; Hume, D.; Impulse response analysis; Mill, J. S.; Monetary policy rules; Observational studies; Potential outcomes; Prediction error variance; Rubin causal model; Simon, H.; Sims non-causality; Structural innovations; Structural vector autoregressions; White noise

JEL Classifications

C32

Granger–Sims causality is based on the fundamental axiom that 'the past and present may cause the future, but the future cannot cause the past' (Granger 1980, p. 330). A variable x then is said to cause a variable y if at time t the variable x_t helps to predict the variable y_{t+1}. While predictability in itself is merely a statement about stochastic dependence, it is precisely the axiomatic imposition of a temporal ordering that allows us to interpret such dependence as a causal connection. The reason is that correlation is a symmetric concept with no indication of a direction of influence, while 'the arrow of time imposes the structure necessary' (Granger 1980, p. 349) to interpret correlations in a causal way.

Definitions

A more precise definition of Granger causality can be given as follows. Assume that all relevant current information is measured in a vector $\chi_t = (y_t, x_t, z_t)$ which is observed at equally spaced

discrete points in time t. (The assumption of discrete time can be relaxed but is maintained here for expositional convenience. Almost all empirical work is in the context of discrete time models.) Denote by Y_s^t all information contained in $y_j, j = \{s,\ldots,t\}$ with equivalent definitions for X_s^t and Z_s^t. (In more formal terms, the expression 'information contained in' can be replaced by 'sigma field generated by'; see Florens and Mouchart 1982, for precise definitions.) Then x_t does not Granger cause y_{t+1} if

$$P\left(y_{t+1} \in A \mid Y_{-\infty}^t, X_{-\infty}^t, Z_{-\infty}^t\right) = P\left(y_{t+1} \in A \mid Y_{-\infty}^t, Z_{-\infty}^t\right) \tag{1}$$

for all t and for any set A for which the conditional probabilities are well defined. It is worth noting that no assumptions of stationarity are needed for this definition. It is common to use the shorthand notation $y_{t+1} \perp X_{-\infty}^t | Y_{-\infty}^t, Z_{-\infty}^t, \forall t$, which states that y_{t+1} and $X_{-\infty}^t$ are independent conditional on $Y_{-\infty}^t$, $Z_{-\infty}^t$. This form of the definition, which imposes a conditional independence restriction on the joint distribution of the process χ_t (see Dawid 1979, for a formal definition and alternative equivalent representations), is similar to the definition used by Granger (1980). It is more general than the original formulations of Granger (1963, 1969), which are based on prediction error variances: Let $E\left(y_{t+1} \mid X_{-\infty}^t\right)$ denote the optimal predictor of y_{t+1} based on $X_{-\infty}^t$, $\varepsilon_{t+1} = y_{t+1} - E\left(y_{t+1} \mid X_{-\infty}^t\right)$ is the prediction error and $\sigma^2\left(y_{t+1} \mid X_{-\infty}^t\right)$ the variance of ε_{t+1}. Then, according to Granger (1969), x_t does not cause y_{t+1} if

$$\sigma^2\left(y_{t+1} \mid Y_{-\infty}^t, X_{-\infty}^t, Z_{-\infty}^t\right) = \sigma^2\left(y_{t+1} \mid Y_{-\infty}^t, Z_{-\infty}^t\right) \tag{2}$$

The conditional independence definition Eq. (1) has the advantage that it does not depend on a particular risk function and is easier to relate to other definitions of causality in stochastic environments such as Suppes (1970) and Rubin (1974). The formulation based on conditional independence was later used and refined in theoretical work by Chamberlain (1982), Florens and Mouchart (1982, 1985), Bouissou et al. (1986) and Holland (1986). The advantage of the prediction error variance definition, Eq. (2) which goes back to Wiener (1956), on the other hand, is that it is easier to implement in statistical tests and has consequently received considerable attention in applied work.

In an influential paper, Sims (1972) showed in the context of covariance stationary processes and restricted to linear predictors that in the case of a bivariate system $\chi_t' = (x_t, y_t)$ the definitions of Granger (1963, 1969) are equivalent to parameter restrictions of the moving average or distributed lag representations of χ_t. When χ_t is covariance stationary it can be represented as

$$\begin{aligned} x_t &= \sum_{j=0}^{\infty} a_j u_{t-j} + \sum_{j=0}^{\infty} b_j v_{t-j} \\ y_t &= \sum_{j=0}^{\infty} c_j u_{t-j} + \sum_{j=0}^{\infty} d_j v_{t-j} \end{aligned} \tag{3}$$

where a_j, b_j, c_j and d_j are constants and u_t and v_t are mutually uncorrelated white noise processes. Sims (1972) shows that the condition 'x_t does not Granger cause $y^t + 1$' is equivalent to c_j or d_j being chosen identically zero for all j. Furthermore, if χ_t has an autoregressive representation, then $x_t = \sum_{j=0}^{\infty} \pi_j y_{t-j} + \varepsilon_t$ where the π_j are parameters and ε_t is an unobservable innovation. Strict exogeneity of y_t in the context of this model is defined as the condition that y_t and ε_s are independent for all values of t and s. Sims then shows that the residuals from a projection of x_t onto $y_{t-j}, j \geq 0$ are uncorrelated with all past and future y_t if an only if x_t does not Granger cause y_{t+1}. In other words, it follows that the condition 'x_t does not Granger cause y_{t+1}' is equivalent to the condition that y_t is strictly exogenous. The relationship between Granger non-causality and strict exogeneity, first discovered by Sims (1972), is further discussed by Engle, Hendry and Richard (1983). Hosoya (1977) shows that this relationship continues to hold in a bivariate setting when processes

are not necessarily stationary and have deterministic components, a situation not originally consider by Granger (1969). The strict exogeneity restrictions discussed in Sims (1972) have become known in the literature as 'Sims noncausality'. In more general terms, and in the context of the process χ_t, the absence of Sims causality of x_t for y_{t+1} is defined as a conditional independence restriction where

$$P\left(x_t \in A \middle| Y_{-\infty}^{\infty}, X_{-\infty}^{t-1}, Z_{-\infty}^{t}\right) = P\left(x_t \in A \middle| Y_{-\infty}^{t}, X_{-\infty}^{t-1}, Z_{-\infty}^{t}\right) \quad (4)$$

for all t or equivalently as $y_{t+1}^{\infty} \perp x_t | Y_{-\infty}^{t}, X_{-\infty}^{t-1}, Z_{-\infty}^{t}, \forall t$. This definition appears in Florens (2003) and Angrist and Kuersteiner (2004) and is closely related to Chamberlain (1982) except that here additional information contained in $Z_{-\infty}^{t}$ is allowed for.

The equivalence between the Granger and Sims notions of causality extends beyond the prediction criterion for covariance stationary processes to definitions of non-causality based on conditional independence restrictions, but, as discussed below, in models with time-varying covariates z_t Sims causality is more appealing and easier to link to the potential-outcomes causality concepts widely used to analyse randomized trials and quasi-experimental studies. Sims (1977, p. 30) emphasizes the use of strict exogeneity restrictions for model identification in structural models and notes that these restrictions can often be related to decision rules of economic agents. Granger non-causality, on the other hand, is a restriction of the reduced form. Structural vector autoregressions (VARs) provide an example of this difference where Sims non-causality imposes restrictions on the impulse response function of structural innovations (Sims 1986, p. 9, discusses the nature of structural innovations and mentions explicitly that they may include random fluctuations of policies) while Granger non-causality imposes restrictions on the reduced form VAR representation of the model. That the two are in general not equivalent will be discussed in Sect. "Equivalence and Non-equivalence of Granger and Sims Causality".

Motivation and Interpretation of Granger–Sims Causality

In order to understand the significance of the original definition of Granger–Sims causality in Granger (1963, 1969) and Sims (1972), it is useful to briefly review the preceding debate in the literature. Simon (1953) defines causality as 'properties of a scientist's model'. Asymmetric functional relationships between variables are given causal interpretations but Simon emphasizes that no notion of time is needed for this definition. In the context of linear systems of equations Simon's definition of causal relationships is equivalent to a block recursive structure of the equations. Simon (1953, p. 65) emphasizes the need for a priori knowledge about the system to identify the block recursive structure. Wold (1954) and Strotz and Wold (1960) strengthen this definition of causality to require a model to be fully recursive. If the model has three variables a, b and c then recursiveness means that a can be solved without knowing b and c, the solution for b generally depends on a, and the solution for c depends on both a and b. Such a system then is interpreted to have a causal relationship where a causes b and a and b both cause c. Wold (1954, p. 166) writes: 'The relationship is then defined as causal if it is theoretically permissible to regard the variables as involved in a fictive controlled experiment.'

The idea is that, because a does not depend on b an c, it can in principle be controlled or changed by an experimenter. In the terminology of Wold (1954, p. 166), a is the cause and b and c are the 'effect variables'. Wold (1954) discusses the distinction between randomized experiments in the sense of R. A. Fisher and nonexperimental observations. In the latter case causal interpretations depend on 'subject-matter theory' (Wold 1954, p. 170) to identify recursive structures. In other words, a priori assumptions about the structure of the model replace randomized experiments as the source of identification. This approach remains popular to this day for the identification of structural VARs where recursive relationships based on a casual appeal to economic theory are imposed. Orcutt (1952) discusses causal chains or triangular

systems but also mentions the possibility of using temporal structures to identify causal links. Basmann (1963) criticizes the recursive identification schemes of Wold and argues in favour of more general structural economic models to identify causal relationships.

Both Granger and Sims voice scepticism about identifying restrictions which can in general not be tested but have to be accepted a priori. Granger (1980, p. 335) writes,

> [i]f these assumptions are correct or can be accepted as being correct, these definitions may have some value. However, if the assumptions are somewhat doubtful, these definitions do not prove to be useful.

And Sims (1972, p. 544; see also 1977, p. 33) notes,

> [i]f one is willing to identify causal ordering with Wold's causal chain form for a multivariate model, and if enough identifying restrictions are available in addition to those specifying the causal chain form, one can test a particular causal ordering as a set of overidentifying restrictions. The conditions allowing such a test are seldom met in practice, however.

The great advantage of Granger's definition of causality is that it is directly testable from observed data. Granger (1969) gives operational definitions and discusses testable parameter restrictions in linear time series specifications. The strict exogeneity restrictions derived by Sims (1972) are particularly revealing of the power of the flow of time as the identifying force behind uncovering causal links between time series. On the assumption that x_t is strictly exogenous, then, if x_t does not cause y_t it must be conditionally independent of future outcomes y_{t+j}. On the other hand, if conditional correlations between future outcomes y_t and current values of x_t are detected, they must be due to a causal influence of x_t on y_t, because, by assumption, the possibility of a causal link between events determining y_{t+j} that lie in the future and current observations of x_t have been excluded and thus cannot be the source of the observed correlation.

As Granger (1963, 1980) notes, the notion of causality has a long and controversial tradition in philosophy. Some treatments discussing relationships to econometric and statistical practice include Pearl (2000) and Hoover (2001). Holland (1986) discusses the causality definitions of David Hume, John Stuart Mill and Suppes (1970) in the context of Rubin's (1974) causal model. Holland points out that Hume's criteria for causality include the axiom of temporal precedence as well as the requirement of a 'constant conjunction' between cause and effect. (See Hoover, 2001, p. 8, for more discussion of Hume's concept of 'constant conjunction'.) Suppes (1970) proposes a probabilistic theory of causality where he replaces constant conjunction with the requirement that

> one event is the cause of another if the appearance of the first event is followed with high probability by the appearance of the second, and there is no third event that we can use to factor out the probability relationship between the first and second events. (Suppes 1970, p. 10)

The definition of Suppes has some parallels to Granger's definition, notably the requirement of temporal succession, the fact that there are no restrictions on what can be a cause and the fact that causes are defined through their effect on the conditional distribution of the effect variable. Finally, Holland attributes the idea of identifying causal effects through experimentation to Mill. Experimentation has since played a central role in the statistical analysis of causality, although Granger (1980, p. 329) mentions it only in passing and does not rely on it in his definition of causality. An important consequence of an experimental concept of causality is that, as Holland (1986, p. 954) writes, 'causes are only those things that could, in principle, be treatments in experiments'. As discussed in Sect. "The Connection Between Granger–Sims Causality and Potential Outcomes", this is a critical difference to the concept of Granger causality which does not restrict possible causes.

Feigl (1953) discusses various aspects of the definition of causality that have appeared historically in philosophy and attempts to extract what he calls a 'purified' definition of causality. Zellner's (1979) critique of the concept of Granger causality is centered on Feigl's definition according to which '[t]he clarified (purified) concept of causation is defined in terms of

predictability according to a law' (Feigl 1953, p. 408). Feigl (1953, p. 417) continues to note that '[p]rediction may be analyzed as a form of deductive inference from inductive premises (laws, hypotheses, theories) with the help of descriptions or existential hypotheses'. Zellner (1979, p.12) writes: 'predictability without a law or set of laws, or as econometricians might put it, without theory, is not causation.' In other words, the causality concept put forward by Feigl is based on a priori theoretical assumptions used to generate predictions, while the Granger–Sims notion of causality replaces these a priori restrictions with the axiom of temporal priority. Feigl (1953, p. 417) notes that causal relationships can be defined even when cause and effect are contemporaneous. According to Feigl, a more important distinction between cause and effect lies in the controllability of the cause as opposed to the effect, which leads Feigl to recommend experimental methods as the best way to identify causal factors. In light of Feigl's work Zellner's main critique of Granger–Sims causality is that it is not based on economic theory to identify causes. (Leamer 1985, also strongly rejects the idea of conducting causal inference without relying on a priori theory.) In a reply to Zellner, Sims (1979, p. 105) notes that Feigl's definition is at least as ambiguous as the term 'causality' itself and that it is so general that it encompasses many other definitions of causality.

Zellner also criticizes three more specific features of Granger causality. First, the requirement that the information set needed to define relevant conditional distributions contain all available information makes the definition non-operational. Zellner (1979, p. 33) writes, 'Granger does not explicitly mention the important role of economic laws in defining the set of "all relevant information"' and emphasizes that additional assumptions beyond statistical criteria are necessary to implement tests for Granger non-causality. Second, the limitation to stochastic phenomena and the assumption or axiom of temporal priority of a cause is unnecessarily restrictive compared with other definitions of causality, such as the one of Feigl (1953), which does not rely on these restrictions. And finally, the use of the prediction error variance as a criterion for predictability and the reliance on an optimal predictor, which according to Zellner may both not be well defined, is too restrictive. As far as this last point is concerned it should be noted that more general definitions of Granger–Sims causality proposed by Granger (1980), Chamberlain (1982) and Florens and Mouchart (1982), which are based on conditional independence restrictions of the joint distribution, do not have the problems that Zellner mentions because these distributional restrictions can be formulated for any process with welldefined conditional distributions.

Zellner further points out that economic theory can play a role in providing overidentifying restrictions that allow directions of causality to be imposed. Sims (1979) objects to this last suggestion on the grounds that a test for causality based on overidentifying restrictions is always a joint test of the correctness of such restrictions and the hypothesis of interest and is thus never conclusive. On the other hand, strict exogeneity and thus Granger non-causality, as pointed out by Sims (1977, p. 30, 33) provides overidentifying restrictions that can be tested for. The scepticism about untestable identifying restrictions is forcefully expressed in Sims (1980).

The role of economic theory in identifying parameters of interest in empirical studies remains one of the most controversial issues in econometrics and empirical economics to date. The debate over the correct definition of causality hinges on what individual researchers are prepared to assume a priori, be it restrictions on the temporal direction of cause and effect or fundamental structures that govern the interaction between economic variables. Granger (1980) does not dispute the potential usefulness of a priori theoretical restrictions in identifying causal relationships but emphasizes the potential for misleading inference should these restrictions turn out to be incorrect.

The problem of specifying the correct information set is recognized by Granger (1969), where it is suggested to restrict the set of all available information to the set of relevant information. Granger (1980) discusses a number of

examples that illustrate the sensitivity of a causal relationship between two variables to additional information in the conditioning set. This problem is also mentioned by Holland (1986). It seems, however, that, even though specification issues are of great importance in applied work, this is not a fundamental limitation of the causality concept put forward by Granger. Moreover, correct specification of relevant conditioning variables is a common problem in most statistical procedures applied to economic data and thus not specific to procedures testing for causality. At the same time, the argument in favour of guidance from economic theory when designing such procedures is probably strongest when it comes to selecting the relevant variables that need to be included in the analysis, a point elaborated in more detail below.

Further problems of interpretation are discussed in Granger (1980). Simultaneity occurs if x_t causes y_{t+1} and y_t causes x_{t+1}. In a bivariate system of equations this form of feedback, as Granger (1969) defines it, typically leads to other inferential problems as discussed in Sims (1972). In particular, the lack of exogeneity in this case invalidates conventional regression methods and complicates the interpretation of reduced form parameters such as in VARs. Furthermore, Granger causality is not a transitive relationship: if x_t causes y_{t+1} and y_t causes z_{t+1} then x_t does not necessarily cause z_{t+1}. Granger (1980, p. 339) gives the following example. Assume that ε_t and η_t are mutually independent i.i.d. sequences and that $x_t = \varepsilon_t$, $y_t = \varepsilon_{t-1} + \eta_t$ and $z_t = \eta_{t-1}$. Then, because $y_t = x_{t-1} + \eta_t$ it is clear that x_t causes y_{t+1}. In the same way, y_t causes z_{t+1} but x_t does not cause z_{t+1} if the conditioning set contains only $X^t_{-\infty}$ and $Z^t_{-\infty}$, which is the typical assumption in bivariate statements of causality. At the same time, x_t does cause z_{t+1} in this example when the information set is enlarged to $(X^t_{-\infty}, Y^t_{-\infty}, Z^t_{-\infty})$ because now the innovation η_{t-1} can be recovered from $\eta_{t-1} = y_{t-1} - x_{t-2}$. This example shows that the concept of Granger–Sims causality can be sensitive to the specification of conditioning information.

Equivalence and Non-equivalence of Granger and Sims Causality

Since the original contributions of Granger (1963, 1969) and Sims (1972), which were mostly cast in terms of forecast error criteria, there has been a sizeable literature concerned with extensions of the basic definition and establishing a number of equivalence relationships. While the conditional independence formulation of Granger causality goes back at least to Granger (1980, p. 330), a formal analysis of the equivalence with a corresponding definition of Sims causality was first obtained by Chamberlain (1982) and Florens and Mouchart (1982). It turns out that the condition for Granger non-causality, which in its more general form can be stated as $y_{t+1} \perp X^t_{-\infty} | Y^t_{-\infty}, \forall t$, does imply the generalized form of Sims non-causality formulated as $y^{\infty}_{t+1} \perp x_t | Y^t_{-\infty}, \forall t$, but the reverse implication does not hold generally. Florens and Mouchart (1982) give a counterexample of a nonlinear process where Sims non-causality holds but Granger non-causality does not hold. As Florens and Mouchart (1982) point out, the two conditional independence relationships are equivalent for Gaussian processes where lack of covariance is equivalent to independence. Chamberlain (1982) shows, on the other hand, that a generalized form of Sims non-causality, stated as $y^{\infty}_{t+1} \perp x_t | Y^t_{-\infty}, X^{t-1}_{-\infty}$, is equivalent to $y_{t+1} \perp X^t_{-\infty} | Y^t_{-\infty}$ under a mild regularity condition limiting temporal dependence. Florens and Mouchart (1982) obtain a very similar result for slightly different definitions of the conditioning sets. General statements of this result can also be found in Bouissou et al. (1986). These authors define additional causality relationships: global non-causality (C) is defined as $y^{\infty}_{t+1} \perp X^t_{-\infty} | Y^t_{-\infty}, \forall t$, Granger non-causality of order k (G_k) is defined as $y^{t+k}_{t+1} \perp X^t_{-\infty} | Y^t_{-\infty}, \forall t$ and Sims non-causality of order k (S_k) is defined as $y^{\infty}_{t+1} \perp X^t_{t-k+1} | Y^t_{-\infty}, X_{t-k}, \forall t$ where X_{t-k} is any subset of $X^{t-k}_{-\infty}$. It is then shown that (G_k), (S_k) and (C) are all equivalent for all k, a result that is also stated in Florens and Mouchart (1982, p. 580). Pierce and Haugh (1977) propose an alternative definition of

Granger causality in the context of linear processes. If $H_{\{Y_0^t, U\}}(y_{t+1})$ is the linear projection of y_{t+1} on $\{Y_0^t, U\}$ where $\{Y_0^t, U\}$ is the closed linear span of all the variables generating Y_0^t and the initial conditions U, then x_t does not Granger cause y_{t+1} if the innovations $y_{t+1} - H_{\{Y_0^t, U\}}(y_{t+1})$ and $x_{p+1} - H_{\{Y_0^p, U\}}(x_{p+1})$ are uncorrelated for all $p \leq t$. Florens and Mouchart (1985) show that this definition is equivalent to covariance-based definitions of Granger and Sims causality under some additional regularity conditions. Generally speaking, the results of this early literature show that Granger causality between two processes x and y is equivalent to appropriate definitions of Sims causality not only in a mean squared prediction error sense but more generally in terms of restrictions on appropriate conditional distributions of the joint process. It should also be noted that these equivalence results continue to hold when both x and y are vector valued processes.

The situation changes, however, quite markedly when an additional set of covariates z is added to the analysis. In this situation Granger non-causality defined as $y_{t+1} \perp X_{-\infty}^t | Y_{-\infty}^t, Z_{-\infty}^t, \forall t$ in general is not equivalent to Sims non-causality defined as $y_{t+1}^{\infty} \perp x_t | Y_{-\infty}^t, X_{\infty}^{t-1}, Z_{-\infty}^t, \forall t$. This result seems to have been first obtained by Dufour and Tessier (1993) for the linear case and also appears in Florens (2003) and Angrist and Kuersteiner (2004) and in the biostatistics literature in Robins et al. (1999). Simple examples can be constructed where x Grangercauses y but does not Sims cause y as well as cases where x Sims causes y but does not Granger cause y. In related work Lütkepohl (1993) and Dufour and Renault (1998) show that, in general, (G_1) does not imply (C) if the information set contains z.

To illustrate the result of Dufour and Tessier (1993), assume that $\chi_t = (y_t, x_t, z_t)$ and that χ_t can be represented as linear functions of present and past structural innovations e_t. To simplify the exposition assume that for $C(L) = A(L)^{-1}$ where $A(L)$ is a matrix of lag polynomials of finite order and L is the lag operator, it holds that $\chi_t = C(L)e_t$. Also assume that the diagonal blocks of $C(0)$, partitioned according to (y_t, x_t, z_t), are full rank. The reduced form VAR representation of χ_t then is $\pi(L)\chi_t = u_t$ with $u_t = C(0)e_t$ and $\pi(L) = C(0)A(L)$. As was discussed before, Sims non-causality imposes zero restrictions on off-diagonal blocks of $C(L)$ while Granger non-causality imposes zero restrictions on corresponding off-diagonal blocks of $\pi(L)$. Now note that when χ_t contains only y_t and x_t, the partitioned inverse formula implies that off-diagonal blocks of $A(L)$ are zero if and only if corresponding blocks of $C(L)$ are zero. Because the latter can hold only if corresponding blocks of $C(0)$ are zero, it follows that $\pi(L)$ has zero off-diagonal blocks if and only if corresponding blocks of $C(L)$ are zero. This is the result of Sims (1972). On the other hand, when $\chi_t = (y_t, x_t, z_t)$ the partitioned inverse formula for matrices partitioned into three blocks shows that $C(L)$ having off-diagonal blocks no longer implies that corresponding blocks of $A(L)$ are necessarily zero as well. Thus the equivalence between Sims and Granger causality no longer holds when additional time varying covariates are included in the analysis. In Sect. "The Connection Between Granger–Sims Causality and Potential Outcomes", applications in monetary economics are discussed where this situation arises naturally.

The connection Between Granger–Sims Causality and Potential Outcomes

The notion of causality that has become standard in micro-econometrics is based on Rubin's (1974) concept of potential outcomes, which at its core uses experimental variation to identify causal relationships. The potential outcomes model has been extended to and applied in observational studies. Observational studies are situations where no experimental assignments of actions were used. Examples are medical trials where experiments might be unethical or many economic policy questions where experiments may be unethical or too expensive to carry out. The importance of experimental evidence was recognized in econometrics dating back to Haavelmo (1944). Wold

(1954) similarly discusses controlled experiments as a way to uncover causal relationships. Orcutt's (1952, p. 305) notion of causality is closely related to the idea of potential outcomes and is defined in terms of consequences of actions:

> Thus when we say that A is the cause of B, we often mean that if A varies, then B will be different in a specified way from what it *would have been* if A had not varied.

Orcutt (1952, p. 309) goes on to discuss policy actions as a substitute for unavailable experimental evidence to identify causal relationships in observational data, an idea explored in more detail below.

For expositional purposes, assume that a certain action or treatment D can be either given or not given to individual i. The causal question in this context is whether the treatment has an effect on an outcome variable of interest measured by y. It is convenient to define $D_i = 1$ if the treatment is given and $D_i = 0$ if the treatment is not given. The potential outcome $y_i(0)$ is defined as the outcome for individual i that would have occurred if the treatment had not been given and $y_i(1)$ as the outcome that would have occurred in the case the treatment had been given. The absence of causality of D for y then is defined as the situation where $y_i(0) = y_i(1)$. This condition is referred to as the 'strong null hypothesis of no causal effect'. Usually this condition cannot be directly tested because $y_i(0)$ and $y_i(1)$ are not both observed for the same individual. Instead, the observed measurement takes the form

$$y_i = D_i y_i(1) + (1 - D_i) y_i(0) \tag{5}$$

Potential outcomes may depend on a list of covariates z_i. Covariates capture characteristics of the outcome variable that are not directly related to the experiment but that need to be taken into account when assessing the outcome. An identification condition is needed to proceed to testable restrictions. Formally one imposes the condition $y_i(0), y_i(1) \perp D_i | z_i$, which is sometimes referred to as selection on observables. This condition is automatically satisfied if D_i is randomly assigned in an experiment. In observational studies the condition essentially states that actions by individuals or policymakers cannot be based on unobservable information. The 'selection on observables' condition implies that

$$P(y_i(j) \in A | z_i) = P(y_i(j) \in A | D_i, z_i) \tag{6}$$

for all j and the null hypothesis implies that

$$P(y_i \in A | D_i, z_i) = P(y_i \in A | z_i) \tag{7}$$

which is identical to Granger's condition of no causal effect, a result that is discussed in Holland (1986). The power of the identifying restriction lies in the fact that it is formulated independently of the null hypothesis of no causal relationship. To be more precise, the identifying restriction imposes conditional independence of D from $y_i(0)$ and $y_i(1)$ but not from y_i. The latter holds only when the null is true. This is an important difference from Sims (1977, p. 30), who writes that 'Causality is an important identifying restriction on dynamic behavioral relations'. In other words, Sims imposes the null to identify certain structural models. Another important difference between this model of causality and the less specific definition of Granger–Sims causality is that the form of the causal link between D and y is of a simple functional form specified in Eq. (5). This particular structure is what allows the interpretation of measured correlations as causal links.

Identification conditions thus lead to testable implications of Rubin's potential outcome framework that are identical to the Granger–Sims definition of non-causality. Nevertheless, causality in Rubin's context is closely related to experiments and counterfactuals: causal effects of a treatment are measured by comparing unobservable counterfactuals under treatment and non-treatment. On the other hand, Granger–Sims causality does not rely on the notion of treatment. It has been applied to studying such phenomena as the temporal link between interest rates and inflation, variables that are endogenously determined and where it is hard to imagine that an experiment or even a policy intervention is available for causal inference. Orcutt's idea of using policy variation can, however, still be implemented if instead of market interest rates one focuses (in the United States)

on the federal funds target rate, a variable that is directly set by the policymaker. Under the additional assumption that all systematic aspects of the policy depend on observable information, it is possible to generate pseudo-experimental variation even in the interest rate example. These ideas are now explored in more detail.

The potential outcomes framework in its original form is in many ways too limited to be directly applicable to macroeconomic questions of causality where the Granger–Sims concept of causality has been mostly applied. The two main limitations of the potential outcomes approach are that it does not allow for dynamic treatments or policies and that usually the stable unit treatment assumption of Rubin (1980) is imposed. The latter rules out general equilibrium effects of treatments and is not satisfied in a macroeconomic context. Angrist and Kuersteiner (2004) propose an extension of the potential outcomes framework that overcomes these limitations (also related is the work of White 2006). Consider an economy that is described by $\chi_t = (y_t, D_t, z_t)$ where y is a vector of outcome variables, D is a vector of policy variables and z is a vector of relevant covariates not already included in y. Potential outcomes $y_{t,j}(d)$ are defined as values the outcome variable $y_{+,j}$ would have taken if at time t the policy had been set to $D_t = d$. It is probably useful to discuss the nature of the potential outcome $y_{t,j}(d)$ in a context where one has a dynamic general equilibrium model describing the evolution of the process χ_t as a system of stochastic difference equations. Then $y_{t,j}(d)$ has to be thought of as a specific solution of that model indexed against a specific decision rule d of some policymaker. It is helpful, but not necessary, to assume that the model has a strong solution, in the sense of stochastic process theory, such that the strong null of no causal effect can be represented as the restriction $y_{t,j}(d) = y_{t,j}(d')$ for all possible values of d. It should be clear from this description that $y_{t,j}(d)$ is a possibly highly complex function of all the inputs that go into the model, including policy decisions taken at times different from t. Solving for $y_{t,j}(d)$ explicitly is not necessary for the definition of a causal link between D and y, a feature that is very much in the spirit of Granger and Sims. All that is needed is an identifying restriction that allows us to interpret observed correlations as causal links. A sufficient condition is

$$y_{t,j}(d) \perp D_t | Y^t_{-\infty}, D^{t-1}_{-\infty}, Z^t_{-\infty}, \forall t, \forall j > 0. \quad (8)$$

Under the sharp null of no causal effect where $y_{t,j}(d) = y_{t+j}$ it then follows immediately that $y_{t+j} \perp D_t | Y^t_{-\infty}, D^{t-1}_{-\infty}, Z^t_{-\infty}$. This is the same as the condition for Sims non-causality. As discussed earlier, it is generally not equivalent to Granger noncausality. The form of the testable restriction depends critically on the form of Eq. (8). At least in cases where D is a decision variable of a policymaker, this restriction leading to Sims causality seems plausible. This can be seen easily in the context of linear models where the identification assumption leading to Sims non-causality is identical to the restriction that policy innovations are independent of all future innovations affecting the outcome variables.

In order to better understand the nature of the identifying assumption Eq. (8) it is useful to consider a specific example. The notion of Granger causality was applied to the question of a causal link between monetary policy and real economic activity, starting with Sims (1972), and thus has a long tradition in the empirical macro literature. Most of the early empirical literature has investigated this question using linear regressions of some measure of monetary aggregates or interest rates and various measures of real economic activity. In an important methodological contribution, Romer and Romer (1989) use information from the minutes of the Federal Open Markets Committee to classify US Federal Reserve policy into times of purely anti-inflationary monetary tightening and other periods. Times of tight monetary policy are called Romer dates. The idea then is to measure average economic activity following Romer dates and to compare these measurements to average economic activity at other times. While the argument that Romer dates are exogenous has been criticized (see for example Hoover and Perez 1994; Shapiro 1994; Leeper 1997), the basic premise of the approach of Romer and Romer (1989) remains valid. It is to use a behavioural theory or policy rule for a policymaker to construct policy innovations which serve as exogenous variation

that can be used to evaluate the effectiveness of the policy in question (Jorda 2005, emphasizes the importance of exogenous variation to identify causal relationships). Angrist and Kuersteiner (2004) analyse the consequences of allowing for additional covariates z in the policy model to capture information about nominal macroeconomic variables such as the inflation rate. These variables are clearly relevant for policy decisions of the Federal Reserve and thus constitute relevant conditioning information in the sense of Granger (1969). At the same time these nominal variables are not part of the null hypothesis of no causal effects on real economic activity and thus cannot be subsumed into the y-process. As discussed earlier, under these circumstances reduced form regressions based on Granger's notion of non-causality cannot be used to test for Sims non-causality.

Generally speaking, a model of the policymaker, in this case the Federal Reserve, is a conditional probability distribution $P\left(D_t \in A \middle| Y^t_{-\infty}, D^{t-1}_{-\infty}, Z^t_{-\infty}\right)$. An example of a policy model for the Federal Reserve in the context of the Romer and Romer data was developed by Shapiro (1994). The fundamental identifying assumption then is that this model is correct, especially in the sense that the conditional probability of D_t does not depend on $y_{t,j}(d)$. This condition will be satisfied when two criteria are met. All relevant information that the policymaker used to decide on the policy D_t is included in the model and the problem at hand is of a nature where the policymaker does not foresee the future. This is the way Granger's fundamental axiom of 'the arrow of time' plays a role to provide identifying assumptions in this setting. If these conditions are met then all random deviations of the policy D_t from what is predicted by the model are conditionally independent of $y_{t,j}(d)$. Random deviations could be due to the variation over time in policymakers' beliefs about the workings of the economy, decision-makers' tastes and goals, political factors, and the temporary pursuit of objectives other than changes in the outcomes of interest (for example, monetary policy that targets exchange rates instead of inflation or unemployment), and finally harder-to-quantify factors such as the mood and character of decision-makers. It is then precisely these random deviations from prescribed policies that help to identify causal links. In other words, it is not the systematic or predictable policy changes that are helpful to answer causal questions but the deviations from prescribed rules. The reason for this is that the causal model used here does not provide enough structure to disentangle causal links from endogenously varying policy variables. The situation is quite similar to the analysis of impulse response functions in structural VAR models where identification is driven by the independence of structural innovations. Impulse response analysis can thus be viewed as a special case of the potential outcomes model when χ_t is a linear process.

The potential outcome framework has the advantage that it focuses on exogenous variation and puts the identification discussion at the centre of causal inference. It helps to clarify the source of identifying variation in an analysis of Granger–Sims causality. A priori arguments for the identifying exogeneity restrictions can be based on institutional settings such as the introduction of new legislation, on procedural details as in the Romer and Romer (1989) example, or on behavioural models derived from economic theory. At the same time the potential outcome approach to identification is limited in the sense that its most natural areas of application lie in the analysis of policy effectiveness. It is less suited to analyse causality between variables that are jointly determined in equilibrium.

A point made by Granger (1980) is also relevant here. The analysis of causality is not necessarily relevant for the analysis of controlled processes. To illustrate the issue, consider the linearized Lucas model of McCallum (1984), where in an overlapping generations framework price setting happens in isolated markets based on local information. McCallum shows that random innovations to the money stock affect unemployment because agents cannot completely distinguish between real price changes in their markets and price changes due to variation in the

supply of money. A test of conditional independence between money and employment for data generated by this model would find evidence of a causal relationship in the sense of Sims. At the same time, any attempt by the monetary authority to systematically exploit this relationship through a systematic policy rule would fail in this model because agents fully incorporate predictable actions of the policymaker and do not respond to nominal changes in prices. This example shows that a statistical definition of causality may indicate the existence of a causal relationship that does not lend itself to policy intervention and control. Whether individual researchers are willing to call such a finding a causal relationship hinges upon their notion of causality and is likely to be controversial.

The situation is reversed in some models where the monetary authority can fully control output through appropriate policy rules. For the purpose of illustration consider the model by Rudebusch and Svensson, which 'consists of an aggregate supply equation that relates inflation to an output gap and an aggregate demand equation that relates output to a short term interest rate' (1999, p. 205). Monetary policy is conducted by setting the nominal interest rate, and affects output and inflation with a one period lag. In this model it is possible for the monetary authority to fully stabilize output such that deviations from a fixed steady state level are serially uncorrelated. On the assumption that the policy rule is augmented with an independently distributed policy innovation (this assumption is necessary for statistical identification of test procedures, see Sims, 1977, p. 39), it follows that a test for Granger causality will not be able to reject the null of no causal relationship running from interest rates to output. At the same time, a test for Granger noncausality of output for interest rates will be rejected because the interest rate setting rule depends on past output. In this example, the direction of causality in the statistical sense of Granger goes in the opposite direction of what the model indicates.

On the other hand, a test of the conditional independence restriction Eq. (4) for Sims non-causality of interest rates for output will be rejected, thus revealing the influence of the policymaker on output. Sims (1977, p. 36) considers a similar reversal of the direction of Granger causality in models where a policy variable is controlled. While Sims considers a bivariate model where both Granger and Sims causality are equivalent, the model of Rudebusch and Svensson considered here has three equations, which explains why the concepts of Granger and Sims causality do not lead to the same conclusions.

A test of Eq. (4) in this model is thus able to identify the direction of causality even when variables are controlled, at least when the test is based on the assumption that the policy model is correctly specified and the policy innovation is thus identified. However, even in this case, the measured causal effects are those of random deviations from the policy rule. As discussed earlier, attempts to exploit these effects with systematic policy actions may not be feasible due to the reactions of rationally forward-looking agents.

A related issue is the problem of analysing causal effects of systematic changes in the policy rule, a problem discussed in Sims (1977, p. 30). Without additional structure such questions seem to be hard to address, and it remains an open question to what extent evidence gained from causal inference based on notions of Granger–Sims causality can be used to investigate them.

Summary

This article explores the notion of Granger–Sims causality as a concept of statistical predictability. The definition is appealing because it does not require a priori theoretical restrictions but rather is formulated in terms of a directly testable implication on the distribution of observed data. The simplicity of this approach to causality has led to extensive applications in areas such as macroeconometrics, where notions of causality that rely on the possibility of carrying out experiments are difficult to apply. Difficulties with a purely statistical concept of causality, however, arise when it comes to interpreting the nature of detected causal relationships. Without additional

assumptions regarding the exogeneity of one or more input variables, it seems difficult to link the statistical causality concept with the more fundamental distinction between cause and effect. The latter distinction is fundamental to the analysis of controllability of outcome variables and thus central to many questions in the social sciences. As discussed above, there is clearly a distinction between a causal link between two variables and the possibility of controlling an output by manipulating certain inputs. Equilibrium effects which are at the core of economic analysis may, for example, pre-empt policy changes through the agent's rational anticipation of just these policy changes. Perceived causal relationships thus may not be exploitable for policy purposes even if they can be reliably identified in the history of an economy. The analysis of causality and controllability in dynamic equilibrium models thus remains a central topic of research.

See Also

- ▶ Causality in Economics and Econometrics
- ▶ Continuous and Discrete Time Models
- ▶ Granger, Clive W. J. (1934–2009)
- ▶ Hume, David (1711–1776)
- ▶ Propensity Score
- ▶ Rubin Causal Model
- ▶ Simon, Herbert A. (1916–2001)
- ▶ Structural Vector Autoregressions

Bibliography

Angrist, J. and G. Kuersteiner. 2004. Semiparametric causality tests using the policy propensity score. Working paper no. 10975. Cambridge, MA: NBER.

Basmann, R. 1963. The causal interpretation of non-triangular systems of economic relations. *Econometrica* 31: 439–448.

Bouissou, M., J.-J. Laffont, and Q. Voung. 1986. Tests of noncausality under Markov assumptions for qualitative panel data. *Econometrica* 54: 395–414.

Chamberlain, G. 1982. The general equivalence of Granger and Sims causality. *Econometrica* 50: 569–581.

Dawid, A. 1979. Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B* 41: 1–31.

Dufour, J.-M., and E. Renault. 1998. Short run and long run causality in time series: theory. *Econometrica* 66: 1099–1125.

Dufour, J.-M., and D. Tessier. 1993. On the relationship between impulse response analysis, innovation accounting and Granger causality. *Economics Letters* 42: 327–333.

Engle, R., D. Hendry, and J. Richard. 1983. Exogeneity. *Econometrica* 51: 277–304.

Feigl, H. 1953. Notes on causality. In *Readings in the philosophy of science*, ed. H. Feigl and M. Brodbeck. New York: Appleton-Century-Crofts, Inc..

Florens, J.-P. 2003. Some technical issues in defining causality. *Journal of Econometrics* 112: 127–128.

Florens, J.-P., and M. Mouchart. 1982. A note on noncausality. *Econometrica* 50: 583–591.

Florens, J.-P., and M. Mouchart. 1985. A linear theory for noncausality. *Econometrica* 53: 157–176.

Granger, C. 1963. Economic processes involving feedback. *Information and Control* 6: 28–48.

Granger, C. 1969. Investigating causal relations by econometric models and crossspectral methods. *Econometrica* 37: 424–438.

Granger, C. 1980. Tests for causation – a personal viewpoint. *Journal of Economic Dynamics and Control* 2: 329–352.

Haavelmo, T. 1944. The probability approach in econometrics. *Econometrica* 12 (Suppl): iii–vi, 1–115.

Holland, P. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81: 945–960.

Hoover, K. 2001. *Causality in macroeconomics*. Cambridge: Cambridge University Press.

Hoover, K., and S. Perez. 1994. Post hoc ergo propter once more: An evaluation of 'does monetary policy matter?' in the spirit of James Tobin. *Journal of Monetary Economics* 34: 47–73.

Hosoya, Y. 1977. On the Granger condition for non-causality. *Econometrica* 45: 1735–1736.

Jorda, O. 2005. Estimation and inference of impulse responses by local projections. *American Economic Review* 95: 162–182.

Leamer, E. 1985. Vector autoregressions for causal inference? *Carnegie-Rochester Conference Series on Public Policy* 22: 255–303.

Leeper, E. 1997. Narrative and VAR approaches to monetary policy: Common identification problems. *Journal of Monetary Economics* 40: 641–657.

Lütkepohl, H. 1993. Testing for causation between two variables in higher dimensional VAR models. In *Studies in applied econometrics*, ed. H. Schneeweiss and K. Zimmerman. Heidelberg: Springer-Verlag.

McCallum, B. 1984. A linearized version of Lucas's neutrality model. *Canadian Journal of Economics* 17: 138–145.

Orcutt, G. 1952. Actions, consequences, and causal relations. *The Review of Economics and Statistics* 34: 305–313.

Pearl, J. 2000. *Causality.* Cambridge: Cambridge University Press.

Pierce, D., and L. Haugh. 1977. Causality in temporal systems. *Journal of Econometrics* 5: 265–293.
Robins, J., S. Greenland, and F. Hu. 1999. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association* 94: 687–700.
Romer, C., and D. Romer. 1989. Does monetary policy matter? A new test in the spirit of Friedman and Schwartz. In *NBER macroeconomics annual 1989*, ed. O. Blanchard and S. Fischer. Cambridge, MA: MIT Press.
Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688–701.
Rubin, D. 1980. Randomization analysis of experimental data: the Fisher randomization test. Comment. *Journal of the American Statistical Association* 75: 591–593.
Rudebusch, G., and L. Svensson. 1999. Policy rules for inflation targeting. In *Monetary policy rules*, ed. J. Taylor. Chicago: University of Chicago Press.
Shapiro, M. 1994. Federal Reserve policy: cause and effect. In *Monetary policy*, ed. G. Mankiew. Chicago: University of Chicago Press.
Simon, H. 1953. Causal ordering and identifiability. In *Studies in econometric method*, ed. W. Hood, T. Koopmans, and Cowles Commission Monograph No. 14. New York: John Wiley.
Sims, C. 1972. Money, income and causality. *American Economic Review* 62: 540–552.
Sims, C. 1977. Exogeneity and causal ordering in macroeconomic models. In *New methods in business cycle research: Proceedings of a conference*, ed. C. Sims. Minneapolis: Federal Reserve Bank of Minneapolis.
Sims, C. 1979. A comment on the papers by Zellner and Schwert. *Carnegie-Rochester Conference Series on Public Policy* 10: 103–108.
Sims, C. 1980. Macroeconomics and reality. *Econometrica* 48: 1–48.
Sims, C. 1986. Are forecasting models usable for policy analysis? *Federal Reserve Bank of Minneapolis Quarterly Review* 10(1): 2–16.
Strotz, R., and H. Wold. 1960. Recursive versus nonrecursive systems: an attempt at synthesis. *Econometrica* 28: 417–427.
Suppes, P. 1970. *A probabilistic theory of causality.* Amsterdam: North-Holland.
White, H. 2006. Time-series estimation of the effects of natural experiments. *Journal of Econometrics* 135: 527–566.
Wiener, N. 1956. The theory of prediction. In *Modern mathematics for the engineer, series 1*, ed. E. Beckenback. New York: McGraw-Hill.
Wold, H. 1954. Causality and econometrics. *Econometrica* 22: 162–177.
Zellner, A. 1979. Causality and econometrics, policy and policy making. *Carnegie-Rochester Conference Series on Public Policy* 10: 9–54.

Graph Theory

Alan Kirman

Abstract

Graphs are used in economics to depict situations in which agents are in direct contact with each other. The use of graph theory enables one to understand the basic properties of the communication network in an economy or market. Typical questions include: how does the structure of a network affect economic outcomes and the welfare of the individuals involved? What happens if agents can choose those with whom they interact? How will networks evolve over time? Theoretical results, economic applications and empirical examples are given.

Keywords

Clusters; Coalitions; Complete and incomplete information; Connectivity; Cores; First order stochastic domination; Graph theory; Matching; Neighbourhoods; Network formation; Networks; Operations research; Power laws; Probability; Small worlds; Spatial economics; Spillover effects; Stochastic graphs; Technological shocks

JEL Classifications

C6

At first sight it might seem that the rather abstract mathematical theory of graphs would be of little relevance for economics. That this is not entirely the case is largely due to developments since the early 1980s. Economists have become interested in the structure of the relations between individuals, firms and groups and their importance for economic activity. These relations can be viewed as a graph, and the properties of the particular graph will have specific implications for the economic outcomes in the situation modelled. A simple example may help. Consider a competitive

economic market. Agents receive prices from a central source and do not communicate with each other. This can be represented as a *star*. The centre which emits the prices is often thought of as the 'Walrasian auctioneer'. However, in many economic situations the organization may be very different, and it is of interest to know what consequences this may have. For example, in an ordinary *n* person game everyone is conscious of what every other player is doing, and this situation could be represented by a complete graph in which there is a link between every pair of players.

While accounts of the use of networks in economics can be found elsewhere (see Goyal 2006; Jackson 2004 for excellent surveys), the question for this article is the extent to which graph theory has helped economists answer the questions they analyse.

Basic Concepts in Graph Theory

First, a few definitions are necessary to set the stage. Think of a set V of *nodes*, (or economic entities), each pair of the nodes will be linked $v(ij) = 1$ or not $v(ij) = 0$, by an *edge*. If there are N nodes then a simple graph can be represented by its *adjacencymatrix* with each element $v(i)j$ being 1 or 0 depending on whether i and j are linked. Obviously, other considerations could be included. For example, if the graph is *directed*, $v(ij) = 1$ does not imply $v(ji) = 1$; an obvious example of this would be an input–output system where a good i enters into the production of j but not vice versa. An *undirected graph*, on the other hand, has a symmetric adjacency matrix, and here a simple example would be that of co-authors of papers.

The *degree*, $d(v)$, of a vertex v is the number of edges with which it is incident. Two vertices are *adjacent* if they are incident to a common edge. The set of *neighbours,* $N(v)$, of a vertex v is the set of vertices which are adjacent to v. The *degree* of a vertex is the number of neighbours it has, or formally, the cardinality of its neighbour set.

A *path* is an alternating sequence of vertices and edges, with each edge being incident to the vertices immediately preceding and succeeding it in the sequence and with no repeated vertices.

The *length* of a path is the number of edges in the sequence defining the walk. If u and v are vertices, the distance from u to v, written $d(u, v)$, is the minimum length of any path from u to v. In an undirected graph, this is obviously a metric. The *diameter, diam*(G) of the graph G is the maximum value of $d(u, v)$, where u and v are allowed to range over all of the vertices of the graph.

A graph is *complete,* ora *clique*, if every pair of vertices are adjacent. The typical n person game can be thought of as involving a *complete graph* since all players are in contact with each other. This is the other extreme from the star of the competitive model.

Economic Applications

With these concepts in mind we can now ask how they may help in analysing economic problems. If we think about situations in which agents or firms are linked through a network, then there are three basic questions. How does the structure of a network affect economic outcomes, which structures are *'efficient'*, and which graph structures or topologies are likely to be found? In a number of applications very different graph structures may govern economic interaction. A first and important distinction is between *exogenous* and *endogenous* graphs. In the former the agents are taken as already linked, and one examines the consequences of the way in which they are linked. One can think of the distinction often made in spatial economics between agents situated on a line or on a lattice. In the latter case the number of neighbours would be four or eight depending on how one constructs the links. In more general networks, the sort of question asked is: how fast will information travel, or how quickly will an epidemic of opinion or innovation grow? The two important characteristics here will be the *connectivity* of a graph and a measure of the maximum distance between two nodes, the *diameter* of the graph.

Another important notion is that of a *neighbourhood* and the burgeoning literature on externalities and spill over effects has come to pay attention to the structure of neighbourhoods and the idea that individuals in certain neighbourhoods may become isolated from the global network (see Durlauf 2004). The important thing here is that neighbourhoods are not necessarily geographical but are determined by the network of interpersonal relations. The architecture of the graph will have important consequences for the allocation of resources and distribution of wealth in an economy.

One can also pose such questions as: how vulnerable is a network to the removal of links? This too is closely tied to the *connectivity* of the graph. This sort of problem has been extensively looked at in the operations research literature, where the vulnerability of a railroad or communications network to a bombing attack was studied over a long period after the Second World War (see, for example, Frank 1967). The same techniques have been applied to the problems of power outages in electricity networks, but this sort of analysis has, for the moment, made few inroads into the economics literature. This problem has also been studied in the context of stochastic networks, as discussed below.

The notion of *vertex degree* is obviously applicable to matching problems and many matching problems can be thought of as finding a bipartite graph satisfying certain efficiency criteria. Wherever there is a limit to the number of agents who can interact, possibly for institutional reasons, this translates into the idea that the degree of a node is constrained.

While these observations give some idea of the use of graph theoretic concepts in economics, there are rather few examples of contributions which draw directly on graph theory. Indeed, it would be fair to say that most of the relevant contributions use concepts from this theory for notational rather than for analytical reasons. An example of a contribution which makes direct appeal to results from graph theory is that of Corominas-Bosch (2004). She looks at buyers and sellers linked by an exogenous network. The outcomes will depend on who is linked to whom, and players choose strategies depending on those chosen by those that they are linked with. She then uses a theorem on the partitions of bipartite graphs in order to analyse the sort of networks of connections that will produce the Walrasian outcome. What is the competitive outcome in this framework? If buyers outnumber sellers, all the surplus goes to the sellers and, conversely, if sellers outnumber buyers all the surplus goes to the buyers. When the numbers are equal the surplus is split 'evenly' between buyers and sellers. She shows that, given certain conditions, either the whole group of players will implement the competitive solution or the group will partition into groups of players in sub-graphs, each of which will implement the competitive solution. In other words, the graph will partition into sub-graphs, each reflecting the overall relationship between buyers and sellers, and the competitive solution will be implemented in each sub-graph. Thus, what Corominas-Bosch shows is the relation between the architecture of the graph of relations and the efficiency of the outcome.

Another example is that of Anderlini and Ianni (1996), who study the long-run properties of learning from neighbours. In their model, at each step the players play a game with one of their m neighbours, who is chosen at random with probability $1/m$. This is repeated and the final result described. Where is graph theory useful here? The idea that one can, in fact, draw at random at each point in time a one-to-one matching of players is simple, but it is not obvious that the structure of the exogenously given links will permit this. The authors wish to limit the structure of relations between individuals to guarantee that it can be done. The situations in which each member of a group is linked to another single member of the same group are considered. Sub-graphs of this type are called 1-*factors*. Obviously, the number of such matchings depends on the underlying graph structure, which is taken as exogenous. Recall that Anderlini and Ianni (1996) require, in the original graph, that the vertex degree of all agents shall be m. However, it is known from graph theory that certain graphs of this sort may permit no such matching, that is, may have no 1-factor. To handle

this, Anderlini and Ianni assume that their graph is the product of m 1-factors, so as to guarantee that they can randomly rematch their players at each stage. Once they can do this they can show whether the learning process used by the agents converges. Here, managing a rather natural, but technical, problem in a rematching game involves graph-theoretic tools.

Much of the literature in economics has focused on *undirected graphs*, although in some situations such as input–output matrices the direction is clearly important, since production processes are not reversible. Evstigneev and Taksar (1995), although working in a stochastic context, have modelled some economic equilibrium problems using *directed graphs* but, as they indicate, the task is made difficult by the fact that some of the mathematical tools which make the undirected case tractable are not available for directed graphs.

Up to this point we have considered the agents as located in a given graph. However, one of the most interesting challenges in examining an economy is to include the evolution of the network structures themselves. If one wants to proceed to a theory of endogenous network formation, a first step might be to find out which organizations of individuals are stable. Thus one would look for 'rest points' of a dynamic process of network evolution. Such rest points would be arrangements, or networks, which would not be subject to endogenous pressures to change them. This, as it stands, is not a well-formulated concept. More of the rules under which agents operate have to be spelled out. The dynamics of the system will be defined by the way in which links are formed. Good surveys of the literature on this subject (see network formation), are available and, once again, although the terms from graph theory are widely used, there is little evidence of the use of graph theoretic results, at least in the deterministic case, to prove economic propositions. In many cases, well-known architectures such as the *star* turn out to be efficient, and it is not difficult to see why. With n agents such a graph has *diameter* 2 and only n–1 links. Thus it is an efficient way of organizing links if links have any costs. The problem is to show that such a structure will emerge as a result of a well-specified interaction process.

Stochastic Graphs

Paradoxically, it is in the case where the economic graph structure is considered to be *stochastic* that more appeal has been made to graph-theoretic results, such as those developed by Erdos and Renyi (1960).

There are different ways of defining a stochastic graph. The simplest one is that used by Erdos and Renyi themselves in which the 1's and the 0's of the adjacency matrix are replaced by probabilities, so that the edge $e(ij)$ exists with probability $p(ij)$, and so the adjacency matrix is now composed of probabilities and if the graph is undirected the matrix will be *symmetric*, that is, $p(ij) = p(ji)$. (This tool was developed in economics by Kirman et al. 1986; Durlauf 1990; Ioannides 1990.) An alternative and more general possibility is to have a probability distribution over all possible adjacency matrices.

An interesting and useful fact is that the graph representing the links through which interaction takes place may become surprisingly highly connected as the number of agents increases, provided that the probability that any two individuals are connected does not go to zero too fast. To understand what is meant by this, consider a result of Bollobas (1985). He shows that, if the probability that any two agents know each other in a graph G with n nodes, $p^n(ij)$ is greater than $\frac{1}{\sqrt{n}}$ for all i and j and for all n, then as n becomes large it becomes certain that the *diameter* of the graph, $D(G^n(p^n(ij)))$ will be at most 2. More formally,

$$\lim_{n\to\infty} \text{Prob}\left(D\left(\Gamma^n\left(p_{ij}^n\right)\right) \leq 2\right) = 1$$

In other words, any two individuals will be sure to have a 'common friend' if the graph is large enough. Thus, as was observed in Kirman et al. (1986), one should say on encountering someone with whom one has a common friend, 'it's a large world'. This somewhat surprising result suggests that, as sociologists have long observed empirically, relational networks are likely to be much more connected than one might imagine.

It was this sort of result which led to the first use of stochastic graphs in economics in Kirman

(1983). The application there was very simple. The *core* of an exchange economy is defined as that set of allocations that no coalition can improve upon, in the sense that no coalition could redistribute its own resources and do better than in the allocation proposed. A classic result, originally due to Edgeworth, shows that, if the set of agents becomes large – that is, we consider a sequence of economies with each economy having more agents than the previous one – then the core shrinks to the competitive equilibria. A standard objection to this is that it is highly implausible that all coalitions should form. The question then is: what may happen if agents communicate only with a certain probability, and only coalitions of agents who are closely linked can form? Suppose that there are n agents and that they can communicate with a fixed probability, so that $p_n(ij) = p_n$. Furthermore, for reasons which are obvious, assume that the probability does not decrease too fast with n, that is $p_n \geq \frac{1}{\sqrt{n}}$. Now we know that it suffices to have large coalitions form for Edgeworth's result to hold. The question that remains is: if n becomes large, will the large coalitions be able to form if we place some restriction on how closely the agents have to be linked to form a coalition? Suppose that we allow coalitions to form only if the maximal distance between two agents (the *diameter* of the sub-graph defined by the coalition) is less than or equal to 2. Now, since the links are drawn at random, the set of allocations is itself a random variable. What we can hope for is that the probability that the core of such a random economy will be different from the set of competitive equilibria converges to zero. As should be clear, it is enough, given our assumptions, to use Bollobas's result directly to obtain this result.

In general, the property of increasing connectivity is of interest, as noted, in economic models, since the connectivity determines how fast information or a 'technological shock' diffuses and how quickly an epidemic of opinion or behaviour will occur. It is important to note that the result just evoked depends crucially on the fact that the actors in the network are linked with uniform probability or, slightly more generally, that the pair of agents with the lowest probability of being linked should still be above the lower bound mentioned.

Emerging Graphs

The dynamic evolution of the state of the individuals linked in a graph-like structure is particularly interesting since the stable configurations of states, if there are any, will depend on the graph in question, and some of the results from other disciplines (see Weisbuch 1990) can be evoked in the context of economic models.

This 'small world' problem discussed above is related to, but rather different from, that studied by Watts (2000), since he looked at networks which evolve as the links are modified stochastically. In other words, he does not consider a realization of a graph drawn from a family of graphs with given probability but, rather, starts with a given graph and shows how structure and, in particular, connectivity may emerge as links are replaced by other links. What he examines is a situation in which agents have a fixed number of links with others, that is, the *vertex degree* of the graph is a constant. For example, they might all be situated on a circle and be linked only with their immediate neighbours. Then one of the existing links is drawn at random and replaced with a new link. This may be to any agent, in particular one to whom the distance was great in the original graph. Adding such links drastically increases the connectivity of the graph. This procedure is repeated and what emerges is a typically clustered structure in which closely linked individuals in a small group are linked to other groups through one or two links.

In a pure random graph of the Erdos-Renyi (E-R) (1960) type, distances are short, as we have seen, but there is almost no clustering. However, in the sort of small world graphs studied by Watts the situation is different: distances are still short but there is a great deal of clustering. Here the few long links between different groups keep the distance down but most of the interaction is within small groups. This sort of clustering is

observed in many empirical situations in economics. In this case the pure random E-R graph is used as a benchmark, but the analysis is self-contained and does not exploit results developed in stochastic graph theory itself.

An example of the direct use of stochastic graph theory concerns the *degree distribution* which, in the case where the set of agents is finite, just specifies the proportion of agents who have each vertex degree k, and which we can denote by $P(k)$. In the E-R case this distribution is Poisson, but more general degree distributions have been considered by Newman et al. (2001), for example. There are two large classes of stochastic graphs. The first is composed of those which have a degree distribution which peaks at a particular k and then declines exponentially for large k. The classic E-R model and the Watts 'small world' model fall into this category. They are relatively homogeneous. The other class is characterized by a degree distribution which decays as a power law, that is,

$$P(k) = k^{-\gamma}.$$

In this class the probability that a node has a very large number of links is much higher than for the previous class. The graphs in this class are referred to as 'scale free' networks.

A first reason for being interested in this division is spelled out by Albert et al. (2000). They show that scale-free networks are very robust to error but very vulnerable to attack. The reason for this is simple. Since in such networks the majority of nodes have a very small number of links, random failure of nodes will have rather little impact. The measure of the impact of the failure of nodes is taken here to be the change in the *diameter* of the graph. However, in E-R graphs, since all nodes have essentially the same number of links, each node contributes equally to the diameter. Thus removing any node will have the same impact. The diameter of the remaining graph will increase progressively and faster than if the same thing is done in the scale-free class. The vulnerability to attack is now clearly in inverse relation to the robustness to errors. Since in the scale-free networks there are nodes with a very large number of links which, as a result, contribute much to the connectivity of the network, an 'enemy' will simply choose one of these nodes as his target and have much more effect than he would on an E-R network. As an illustration, Albert et al. (2000) show that the removal at random of 2.5% of the nodes in the World Wide Web, whose degree distribution is well fitted by a power law (see Faloutsos et al. 1999), has no significant effect on its diameter. However, the removal of 2.5% of the most connected nodes in an E-R graph increases the diameter by a factor of 6.

The *degree distribution* is of interest for other reasons. It permits one to work out the average number of neighbours at *distance m* from the agents in a graph. This number can be calculated for each m and, if this sequence converges, there cannot be a *giant component* and all the agents will belong to small connected components in the graph. If this number diverges, however, there will always be a giant component containing many agents and smaller connected sets. There is a crucial level of the ratio of the average number of neighbours at distance 1 to the average number at distance 2 which determines which of these situations applies. This value will depend on the *degree distribution*. Such threshold values play an important role in Erdos and Renyi's work. Newman's interest has tended to focus on networks of co-authors and citations, and the obvious interest here is the extent to which there is a large disciplinary, or interdisciplinary, network, and the extent to which there are fragmented self-referential networks. It is clear, however, that the same ideas could be applied to trading groups or to firms and sectors.

From the empirical evidence one can form an estimation of the degree distribution in a particular network, and one can then examine the distribution of different sized components and pose the question as to the relation between the two and its consistency with the asymptotic predictions.

The degree distribution has been used in another context by Galeotti et al. (2006). They

study the results for games played on networks and where payoffs depend on the players' own actions and those of their neighbours, and examine the influence of three features (whether the games involve strategic substitutes or complements, negative or positive externalities, and incomplete or complete information). It is the latter aspect that is of interest here. They say that an agent has *incomplete information* if a player knows only his own degree and the degree distribution, and *complete information* if he knows the degree of every player. Under incomplete information they show, for example, in games with positive, (negative) externalities, expected payoffs from a game are increasing (decreasing) in degree. This is in contrast with earlier results of Bramoullé and Kranton (2007), where in a public goods game players with higher degrees earn worse payoffs. The key difference from the result just mentioned is that the implicit assumption in their paper is that there is *complete information.*

The assumptions underlying the model developed by Galeotti et al. (2006) are that the players all believe that the graph in which they are situated is drawn from a family of networks that has two properties:

(i) the probability of any node having k links is $P(k)$; and
(ii) the degrees $k_i(g)$ and $k_j(g)$ of any two players i and j are stochastically independent.

This is interesting since it harks back to the 'configuration' model described in Bollobas (1985). (Ioannides (2006), describes the case where the degrees of neighbouring agents are no longer independent, and develops the theory of Markov random graphs, which allow for dependence between neighbouring nodes.)

Another idea developed by Galleotti et al. (2006) is that of looking at the results for games played on different graphs, in particular where the degree distribution of one is more or less 'spread out' than the other. Slightly more formally, one looks at whether one distribution first order stochastically dominates (FOSD) the other or vice versa. In such cases for particular games, they are able to show a relationship between the payoff for a player with a given vertex degree when a particular type of game is played on a graph g and the payoff when the game is played on another graph g'. If g FOSD another graph g', then the payoff is higher under g than under g'. Unfortunately, no such unambiguous results are available under more general changes in the degree distribution.

Conclusion

Economists are often accused of the wholesale borrowing of results from various branches of mathematics and using them even in inappropriate contexts. The growing literature on the importance of networks in economics seems to provide a counter-example. Although many of the concepts and even the notation of mathematical graph theory are used, rather little of the formal structure has been imported, and most of that has been concentrated on stochastic graphs. Many results in the economic literature have, however, been proved *ex nihilo*, and one could argue that economists have in this modest way added to the graph theory literature.

See Also

- ▶ Externalities
- ▶ Network Formation
- ▶ Network Goods (Empirical Studies)
- ▶ Network Goods (Theory)

Bibliography

Albert, R., H. Jeong, and A.-L. Barabasi. 2000. Error and attack tolerance of complex networks. *Nature* 406: 378–382.

Anderlini, L., and A. Ianni. 1996. Path dependence and learning from neighbours. *Games and Economic Behavior* 13: 141–177.

Bollobas, B. 1985. *Random graphs*. London: Academic Press.

Bramoullé, Y., and R. Kranton. 2007. Public goods in networks. *Journal of Economic Theory* 135: 478–494.
Corominas Bosch, M. 2004. Bargaining in a network of buyers and sellers. *Journal of Economic Theory* 115: 35–77.
Durlauf, S.N. 1990. *Locally interacting systems, coordination failure, and the behavior of aggregate activity.* Working paper, Departmetn of Economics, Stanford University.
Durlauf, S.N. 2004. Neighborhood effects. In *Handbook of regional and urban economics*, ed. J.V. Henderson and J.F. Thisse. Elsevier: North-Holland.
Erdos, P., and A. Renyi. 1960. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5: 17–60.
Evstigneev, I.V., and M. Taksar. 1995. Stochastic equilibria on graphs. *Journal of Mathematical Economics* 24: 383–406.
Faloutsos, M., P. Faloutsos, and C. Faloutsos. 1999. On power-law relationships of the internet topology, ACM SIGCOMM '99. *Computers and Communication Review* 29: 251–263.
Frank, H. 1967. Vulnerability of communication networks. *IEEE Transactions on Communications* 15: 778–779.
Galeotti, A., S. Goyal, M. Jackson, F. Vega-Redondo, and L. Yariv. 2006. *Network games*. Mimeo: University of Essex and California Institute of Technology.
Goyal, S. 2006. Learning in games. In *Group formation in economics, networks, clubs and coalitions*, ed. G. Demange and M. Wooders. Cambridge, UK: Cambridge University Press.
Ioannides, Y.M. 1990. Trading uncertainty and market form. *International Economic Review* 31: 619–638.
Ioannides, Y.M. 1997. Evolution of trading structures. In *The economy as an evolving complex system II*, ed. W.B. Arthur, S.N. Durlauf, and D.A. Lane. Reading: Addison Wesley.
Ioannides, Y.M. 2006. *Random graphs and social networks*. Discussion Paper No. 518, Department of Economics, Tufts University.
Jackson, M. 2004. A survey of models of network formation: Stability and efficiency. In *Group formation in economics; networks, clubs and coalitions*, ed. G. Demange and M. Wooders. Cambridge, UK: Cambridge University Press.
Kirman, A. 1983. Communication in markets: A suggested approach. *Economics Letters* 12: 101–108.
Kirman, A.P., C. Oddou, and S. Weber. 1986. Stochastic communication and coalition formation. *Econometrica* 54: 129–138.
Newman, M., S.H. Strogatz, and D.J. Watts. 2001. Random graphs with arbitrary degree distributions and their applications. *Physical Review E* 64: 026118.
Watts, D. 2000. *Small worlds*. Princeton: Princeton University Press.
Weisbuch, G. 1990. *Complex system dynamics*. Redwood City: Addison-Wesley.

Graphical Games

Michael Kearns

Abstract

Graphical games and related models provide network or graph-theoretic means of succinctly representing strategic interaction among a large population of players. Such models can often have significant algorithmic benefits, as in the NashProp algorithm for computing equilibria. In addition, several studies have established relationships between the topological structure of the underlying network and properties of various outcomes. These include a close relationship between the correlated equilibria of a graphical game and Markov network models for their representation, results establishing when evolutionary stable strategies are preserved in a network setting, and a precise combinatorial characterization of wealth variation in a simple bipartite exchange economy.

Keywords

Computation of Equilibria; Correlated Equilibrium; Dynamic Programming; Evolutionary Game Theory; Evolutionary Stable Strategies; Exchange Economies; Graphical Economics; Graphical Games; Network Structure

JEL Classifications

C7

Graphical games are a general parametric model for multi-player games that is most appropriate for settings in which not all players directly influence the payoffs of all others, but rather there is some notion of 'locality' to the direct strategic interactions. These interactions are represented as an undirected graph or network, where we assume that each player is identified with a vertex, and that the payoff of a given player is a function of only his or her own action and those of his or her

immediate neighbours in the network. Specification of a graphical game thus consists of the graph or network, along with the local payoff function for each player. Graphical games offer a number of representational and algorithmic advantages over the normal form, have permitted the development of a theory relating network topology to equilibrium properties, and have played a central role in recent results on the computational complexity of computing Nash equilibria. They have also been generalized to exchange economies, evolutionary game theory, and other strategic settings.

Definitions

A graphical game begins with an undirected graph or network $G = (V, E)$, where V is the set of players or vertices, and E is a set of edges or unordered pairs of vertices/ players. The assumed semantics of this graph are that the payoffs of players are determined only by their local neighbourhoods. More precisely, if we define the neighbour set of a player u as $N(u) = \{v: (u, v)$ is in $E\}$, the payoff of u is assumed to be a function not of the joint action of the entire population of players, but only the actions of u itself and the players in $N(u)$. Complete specification of a graphical game thus consists of the graph G, and the local payoff functions for each player. Note that at equilibrium, it remains the case that the strategy of a player may be indirectly influenced by players arbitrarily distant in the network; it is simply that such influences are effected by the propagation of the local, direct payoff influences.

In the case that G is the complete network, in which all pairs of vertices have an edge between them, the graphical game simply reverts to the multi-player normal form. However, in the interesting cases the graph may exhibit considerable asymmetry and structure, and also be much more succinct than the normal form. For instance, if $|N(u)| <= d$ for all players u, then the total number of parameters of the graphical game grows exponentially only in the degree bound d, as opposed to exponentially in n for the normal form. Thus when d is much smaller than n (a reasonable expectation in a large-population game with only local interactions), the graphical game representation is exponentially more parsimonious than the normal form. Qualitatively, one can think of graphical games as a good model for games in which there may be many players, but each player may be directly and strongly influenced by only a small number of others. Graphical games should be contrasted with other parametric models such as congestion and potential games, in which each player has global influence, but often of a highly specific and weak form. They can also be viewed as a natural generalization of more specific network-based games studied in the game theory and economics literature (Jackson 2007).

Computational Properties

In addition to the aforementioned potential for representational parsimony, graphical games permit a family of natural and sometimes provably efficient algorithms for the computation of Nash and other equilibria. It should be emphasized here that by 'efficient' we mean an algorithm whose running time is a 'slowly' growing function of *the number of parameters of the graphical game representation* (which may be considerably more challenging than a slowly growing function of the number of parameters in the normal form representation, which may be much larger). As is standard in computer science, 'slowly' growing typically means a polynomial function (ideally of low degree).

For instance, in the special case that the graph structure is a tree (or can be modified to a tree via a small number of standard topological operations involving, for instance, the merging of vertices), there is an algorithm running in time polynomial in the number of parameters that computes approximations to (all) Nash equilibria of the given graphical game (Kearns et al. 2001.) This algorithm is based on dynamic programming, and is decentralized in the sense that communication need take place only between neighbouring vertices in the network. In even more restrictive topologies, efficient algorithms for computing exact Nash equilibria are known (Elkind et al. 2006). For non-tree topologies, a generalization of this algorithm known as NashProp (Ortiz and Kearns

2002) has been developed that is provably convergent, but has weaker guarantees of computational efficiency. Provably efficient algorithms have also been developed for computing correlated equilibria (again in the sense of the computation time being polynomial in the number of parameters) for general graphical games (Papadimitriou 2005; see also Kakade et al. 2003). These algorithmic results are in sharp contrast to the status of computing equilibria for games represented in the normal form, where the results are either negative or remain unresolved. Graphical games have also proven valuable in establishing computational barriers to computing Nash equilibria efficiently, and certain classes of graphical games have been shown to be just as hard as the normal form in this regard (Daskalakis and Papadimitriou 2005, 2006; Daskalakis et al. 2006; Schoenebeck and Vadhan 2006).

Extensions of the Model

Since the introduction of graphical games, a number of related models have been introduced and studied. In each case, the model again begins with an undirected network in which the edges represent the pairs of participants that are permitted to interact directly in some strategic or economic setting.

For instance, the model known as *graphical economics* (Kakade et al. 2004) provides a network generalization of the classical exchange economies studied by Arrow and Debreu and others. As in the classical models, each consumer has an initial endowment over k commodities, and a subjective utility function describing his or her preferences over bundles of commodities. However, unlike the classical model, not all pairs of consumers may engage in trade. Instead, each consumer or vertex u may only trade with his or her neighbours $N(u)$, and there is no resale permitted. Equilibria in prices and consumption plans can still be shown to always exist, but now the equilibrium prices may need to be *local*, in the sense that two consumers may charge different prices per unit for the same commodity, and these prices may depend strongly on the network topology. This introduces variation in equilibrium wealth dependent on a consumer's position in the overall network (see below). As with graphical games, the graphical economics model permits efficient computation of equilibria under certain topological restrictions on the network.

More recently, a network version of evolutionary game theory (EGT) has also been examined (Kearns and Suri 2006). In classical EGT, there are random encounters between all pairs of organisms; in the network generalization, such encounters are restricted to the edges of an undirected network. Thus the evolutionary fitness of an organism represented by vertex u is once again determined only by the strategies of its neighbours $N(u)$. More than one reasonable generalization of the evolutionary stable strategy (ESS) of classical EGT is possible in the network setting.

As with graphical games, both graphical economics and the network EGT model revert to their classical counterparts in the special case of the complete graph over all participants, and thus represent strict generalizations, but which are most interesting in cases where the underlying graph has some non-trivial structure.

Network Structure and Equilibrium Properties

Aside from the algorithmic properties discussed above, one of the most interesting aspects of the various models under discussion is their ability to permit the study of how equilibrium properties are influenced by the structure or topology of the underlying network, and indeed there is a growing body of results in this direction.

In the case of graphical games, a tight connection can be drawn between the topology of the underlying graph G of the game and the structure of any minimal (in a certain natural technical sense) correlated equilibrium (CE), regardless of the details of the local payoff functions. Among other consequences, this result implies that CE in graphical games can be implemented using only local, distributed

sources of randomization throughout the network in order effect the needed coordination, rather than the centralized randomization of classical CE. The result also provides broad conditions under which the play of two 'distant' players in the network may be conditionally independent – for instance, at CE, the play of vertices u and v is always conditionally independent given the pure strategies of any vertex cutset between them (Kakade et al. 2003).

In the graphical economics model, for certain simple cases one can give precise relationships between equilibrium wealth, price variation and network structure. For instance, in the case of an arbitrary bipartite network for a simple two-commodity buyer–seller economy with symmetric endowments and utilities (thus deliberately rendering network position the only source of asymmetry between consumers), there is no price or wealth variation at equilibrium if and only if the underlying network has a perfect matching sub-graph between buyers and sellers. More generally, a purely structural property of the network characterizes the ratio between the greatest and least consumer wealth at equilibrium. These structural results have been applied to analyse price and wealth variation in certain probabilistic models of buyer–seller network formation. For instance, it has been shown that whereas truly random networks with a certain minimum number of edges generally exhibit no variation in prices or wealth, those generated by recent models of social network formation such as preferential attachment lead to a power-law distribution of wealth (Kakade et al. 2004).

In the networked EGT setting, it has been proven that even networks with rather sparse connectivity (in which each organism directly interacts with only a small fraction of the total population), but in which the connections are formed randomly, classical ESS are always preserved, even if the initial locations of the invading population are arbitrary. Alternatively, if the network is arbitrary but the initial locations of the invading population are selected randomly, classical ESS are again preserved (Kearns and Suri 2006). Related network models include those of Blume (1995) and Ellison (1993).

See Also

- ▶ Computation of General Equilibria
- ▶ Learning and Evolution in Games: Ess
- ▶ Mathematics of Networks
- ▶ Stochastic Adaptive Dynamics

Bibliography

Blume, L.E. 1995. The statistical mechanics of best-response strategy revision. *Games and Economic Behavior* 11: 111–145.

Daskalakis, C., and C. Papadimitriou. 2005. Computing pure Nash equilibria via Markov random fields. *Proceedings of the 6th ACM conference on electronic commerce*.

Daskalakis, C., and C. Papadimitriou. 2006. The complexity of games on highly regular graphs. The 13th annual European symposium on algorithms.

Daskalakis, C., P.W. Goldberg, and C.H. Papadimitriou. 2006. The complexity of computing a Nash equilibrium. *Proceedings of the 38th ACM symposium on theory of computing*.

Elkind, E., L.A. Goldberg, and P.W. Goldberg. 2006. Nash equilibria in graphical games on trees revisited. *Proceedings of the 7th ACM conference on electronic commerce*.

Ellison, G. 1993. Learning, local interaction, and coordination. *Econometrica* 61: 1047–1071.

Jackson, M. 2007. The study of social networks ineconomics. In *The missing links: Formation and decay of economic networks*, ed. J. Podolny and J.E. Rauch. New York: Russell Sage Foundation.

Kakade, S., M. Kearns, J. Langford, and L. Ortiz. 2003. Correlated equilibria in graphical games. *Proceedings of the 4th ACM conference on electronic commerce*.

Kakade, S., M. Kearns, L. Ortiz, R. Pemantle, and S. Suri. 2004a. Economic properties of social networks. *Neural information processing systems 18*.

Kakade, S., M. Kearns, and L. Ortiz. 2004b. Graphical economics. *Proceedings of the 17th conference on computational learning theory.*

Kearns, M., and S. Suri. 2006. Networks preserving evolutionary stability and the power of randomization. *Proceedings of the 7th ACM conference on electronic commerce*.

Kearns, M., M. Littman, and S. Singh. 2001. Graphical models for game theory. *Proceedings of the 17th conference uncertainty in artificial intelligence*.

Ortiz, L., and M. Kearns. 2002. Nash propagation for loopy graphical games. *Neural information systems processing 16*.

Papadimitriou, C.. 2005. Computing correlated equilibria in multi-player games. *Proceedings of the 37th ACM symposium on the theory of computing*.

Schoenebeck, G., and S. Vadhan. 2006. The computational complexity of Nash equilibria in concisely represented games. *Proceedings of the 7th ACM conference on electronic commerce*.

Gras, Norman Scott Brien (1884–1956)

Thomas C. Cochran

Gras was born in 1884 in Toronto, Canada. His family were not well off, but high intelligence won him scholarships and made possible his education, first at the University of Western Ontario, and then at Harvard with a PhD thesis directed by Edwin F. Gay. He taught at Clark University in Worcester, Massachusetts, and then from 1918 to 1927 at the University of Minnesota, where he was Professor of English History.

In 1927 he was invited to fill the new Strauss Chair in History at the Harvard Business School. Although a considerable amount of American work had been done in business history since the 1880s, Gras was the first holder of an endowed chair. He took as his field the study of business in the history of capitalism and adopted time-based subdivisions of 'petty', 'mercantile', 'industrial', 'financial' and 'national'. In the later periods, particularly, the study of the business firm became the dominant theme.

A *Journal of Economic and Business History* started under the editorship of Gay in 1928, had to be abandoned for lack of funds in 1932. In 1938, Gras and local business friends organized the Business Historical Society with a semi-annual *Bulletin* that led in 1954 to the Quarterly *Business History Review.* The depression was probably responsible to some degree for a narrowing of Gras's own emphasis and that of his students to the history of the firm, usually a self-sustaining type of publication.

In 1939 Gras published his *Business and Capitalism*: An *Introduction to Business History*, which illustrates his historical rather than theoretical emphasis. Two years before his retirement in 1950s, Gras and Professor Henrietta M. Larson of the Business School founded the Business History Foundation to write the history of Standard Oil and other companies that might apply. Gras died in 1956 at Cambridge, Massachusetts.

Selected Works

1915. *Evolution of the English corn market from the twelfth to the eighteenth century*, Harvard economic studies, vol. 13. Cambridge, MA: Harvard University Press.

1918. *The early English customs system: A documentary study of the institutional and economic history of the customs from the thirteenth to the sixteenth century*, Harvard economic studies, vol. 18. Cambridge, MA: Harvard University Press.

1922. *An introduction to economic history.* New York: Harper.

1925. *A history of agriculture in Europe and America*, 2nd ed. New York: Crofts, 1940.

1930a. *Industrial evolution*. Cambridge, MA: Harvard University Press.

1930b. (With Ethel C. Gras.) *The economic and social history of an English village (Crawley, Hampshire), AD 909–1928*, Harvard economic studies, vol. 34. Cambridge, MA: Harvard University Press.

1937. *The Massachusetts first national bank of Boston: 1784–1934*, Harvard studies in business history, vol. 4. Cambridge, MA: Harvard University Press.

1939a. *Business and capitalism: An introduction to business history.* New York: Crofts.

1939b. (With Henrietta M. Larson.) *Casebook in American business history.* New York: Crofts.

1942. *Harvard cooperative society past and present: 1882–1942.* Cambridge, MA: Harvard University Press.

1962. In *Development of business history up to 1950*, ed. Ethel C. Gras. Lincoln Educational Foundation. (Published posthumously.)

Graunt, John (1620–1674)

R. M. Smith

Keywords

Bills of Mortality; Fertility; Graunt, J.; Life tables; Mortality; Petty, W.; Population growth

JEL Classifications

B31

Graunt was born on 24 April 1620 in Hampshire and died on 18 April 1674 in London. At the age of 16 he was apprenticed to his father as a haberdasher of small wares, but remarkably little is known of his life before he published his *Natural and Political Observations Made upon the Bills of Mortality* in 1662. Graunt had formed a friendship with William Petty, who came from a social and economic background similar to his own and who may have drawn Graunt's attention to the data in the London Bills of Mortality. Six months after the publication of the *Natural and Political Observations* Graunt was made a fellow of the Royal Society, in whose foundation Petty had been greatly involved. The publication of this volume had an immediate impact; a second edition was published the same year and two others in 1665. Graunt subsequently fell into disgrace following his conversion to Catholicism and he died in poverty despite generous help from Petty. The latter's own work, especially that on urban growth, owed much to methodologies initiated by Graunt.

It has been claimed that the *Natural and Political Observations* 'created the subject of demography' (Glass 1963) as it involved the first truly analytic study of births and deaths within a population precisely situated in space and time. To do this Graunt employed the Bills of Mortality (Greenwood 1948) and the records of christenings in 17th-century London in order to investigate mortality and population growth in the city. The study's most outstanding qualities are revealed by the search for regularities and configurations in mortality and fertility along with a critical and very insightful appreciation of the quality of the data. He was greatly concerned to establish mortality rates by age and through this interest he came remarkably close to constructing the first formal life table. Using the cause of death evidence in the Bills of Mortality, Graunt without any information whatsoever on ages proceeded to estimate the extent of mortality in infancy and childhood by selecting those causes of death which he guessed would only affect children 'under four or five years old'. To these he added half of the deaths from smallpox and measles which he thought would fall upon children under six, along with slightly less than one third of the plague victims. From these assumptions he derived an estimate suggesting that 36 per cent of all deaths occurred to children under six years. Similar principles were employed to estimate that seven per cent of deaths through 'ageing' could be attributed to those over 70 years. He then proceeded to take these not implausible estimates of mortality at young and old ages to construct an elementary life table which unfortunately was flawed by the unrealistic assumption that above age six the deaths in each specified age period amounted to about three-eighths of the survivors at the beginning of the period. Nonetheless, what Graunt had evolved was an outstanding innovation and was very soon taken up and developed by others.

Graunt's interests in the *Natural and Political Observations* were innovative in other respects; he attempted to calculate the size and rate of growth of 17th-century London, the incidence of plague mortality (Sutherland 1963, 1972), and sex ratios at birth and levels of maternal mortality, and he made some highly believable estimates of the levels of immigration to London that were needed to sustain the city's remarkable growth in the 17th century.

Selected Works

1662. *Natural and political observations . . . upon the Bills of Mortality.* There was a further edition later in 1662, and the plague of 1665 called forth a third edition in the early summer and a

fourth edition in November (printed in Oxford) of that year.
1938. *Natural and political observations,* ed. W.-F. Wilcox. Baltimore: Johns Hopkins University Press.
1973. Natural and political observations made upon the Bills of Mortality. In *The earliest classics,* ed. P. Laslett. Farnborough: Gregg International Publishers.

Bibliography

Glass, D.V. 1963. John Graunt and his 'Natural and political observations'. *Royal Society of London Notes and Records* 19: 63–100.
Greenwood, M. 1948. *Medical statistics from Graunt to Farr*. Cambridge: Cambridge University Press.
Sutherland, I. 1963. John Graunt: A tercentenary tribute. *Proceedings of the Royal Society, Series A* 126: 537–556.
Sutherland, I. 1972. When was the Great Plague? In *Population and social change*, ed. D.V. Glass and R. Revelle. London: Edward Arnold.

Gravity Equation

Robert C. Feenstra

Abstract

The gravity equation explains the amount of trade between countries based on their economic sizes and the distance between them. While it has been in use since the 1960s, its theoretical foundation has been known for a much shorter period, and recent years have seen an large amount of research on its derivation and estimation. We review the theoretical and empirical literature on the gravity equation. In addition to explaining the amount of trade, this equation has been applied to foreign direct investment, the volatility of prices, and the impact of currency unions and free trade areas.

Keywords

Border effects; Constant-elasticity-of-substitution (CES) preferences; Currency unions; Distance; Elasticity of substitution; Estimation; Euro; Fixed effects; Foreign direct investment; Gravity equation; International trade; Monopolistic competition; Poisson distribution; Product differentiation; Selection equation; Specialization; Transport costs; World Trade Organization

JEL Classifications

F

The importance of countries' economic size in explaining trade patterns was recognized in an equation first proposed by Tinbergen (1962). Tinbergen proposed that the volume of trade between countries would be similar to the force of gravity between objects. Suppose that two objects each have mass M_1 and M_2, and they are located distance d apart. Then, according to Newton's universal law of gravitation, the force of gravity F_g between these two objects is:

$$F_g = G = \frac{M_1 M_2}{d^2},$$

where G is the gravitational constant, M_1 and M_2 the mass of the two objects, and d the distance between them. The larger each of the objects are, or the closer that they are to each other, then the greater is the force of gravity between them.

The equation proposed by Tinbergen to explain trade between countries is very similar to Newton's law of gravity, except that instead of the mass of two objects we use the gross domestic product (GDP) of two countries, and instead of predicting the force of gravity we are predicting the amount of trade between them. So the gravity equation in trade is:

$$\text{Trade} = A\frac{\text{GDP}_1\text{GDP}_2}{\text{d}^{\gamma}}, \qquad (1)$$

where trade is the amount of trade (that is, imports, exports, or their sum total) between two countries, GDP_1 and GDP_2 their gross domestic products, d the distance between them, and A a constant. We use the exponent γ on d^{γ} rather than

d^2 as in Newton's law of gravity, because we are not sure of the economic relationship between distance and trade.

While Tinbergen (1962) showed that the gravity equation worked well empirically, it was some years before the theoretical foundation of this equation was known. The earliest contribution is Anderson (1979), followed by Bergstrand (1985, 1989) and Helpman (1987). All these authors suppose that *countries produce different goods*. That result might be due to the underlying assumption of monopolistic competition, in which case every firm produces a different product variety from every other firm. It follows that firms in different countries also produce different product varieties (which we label as different goods). But that is not the only model where countries produce different goods. That result also applies in a perfectly competitive model where there are more goods than factors, and some differences in factor prices or technologies across countries (Bhagwati 1972; Davis 1995).

Along with country specialization in different goods, let us add the assumption that the utility function of the representative consumer is homothetic, so that the income elasticities of demand are unity (that is, the share of demand spent on each good does not vary with income). In that case, we can follow Helpman (1987) and derive a simple gravity equation. Let $i,j = 1, \ldots, C$ denotes countries, and let $k = 1, \ldots, N$ denotes goods. Let y_k^i denote country i's production of good k. Assume that prices are the same across all countries due to free trade and no transport costs, and normalize the prices at unity, so y_k^i actually measures the *value* of production. Total GDP in each country is $Y^i = \sum_{k=1}^{N} y_k^i$, and world GDP is $Y^w = \sum_{i=1}^{C} y^i$.

Let s^j denote country j's share of world expenditure. If we assume that trade is balanced in each country, then s^j also denotes country j's share of world GDP, so that $s^j = Y^j/Y^w$. Then with all countries producing different goods, and with identical and homothetic demand across countries, the *exports from country i to county j of product k* are given by:

$$X_k^{ij} = s^j y_k^i. \tag{2}$$

Summing over all products k, we obtain:

$$X^{ij} = \sum_k X_k^{ij} = s^j \sum_k y_k^i = s^j Y^i = \frac{Y^j Y^i}{Y^w}. \tag{3}$$

So this simple model gives us a gravity equation like (1), where $A = 1/Y^w$. The gravity equation in (3) omits a distance term, however. Notice that, if instead we solve for X^{ji} then we get $Y^j T^i/Y^w$, which is exactly the same as (3). So in this very simple model, bilateral exports equal bilateral imports (that is, trade is balanced between every pair of countries).

The key limitation of the gravity Eq. (3) is that it assumes no transport costs. Since transport costs will depend on the distance between countries, by introducing them we will also introduce a distance term into the gravity equation. But allowing for transport costs means that the prices of goods will differ across countries, since a distant country will have higher c.i.f. prices (that is, including cost, insurance and freight charges). In that case we need to re-derive the gravity equation while allowing for different prices. To do so requires us to introduce an explicit demand structure into the model. A commonly used demand within the monopolistic competition model arises from constant-elasticity-of-substitution (CES) preferences. In that case the demand for a good i exported by that country-to-country j is:

$$c^{ij} = \left(p^{ij}/P^j\right)^{-s}\left(Y^j/P^j\right), \tag{4}$$

where $\sigma > 1$ is the elasticity of substitution, p^{ij} the c.i.f. price of the good exported from i to j, and P^j refers to country j's overall price index, defined as:

$$P^j = \left(\sum_{i=1}^{C} N^i \left(p^{ij}\right)^{1-\sigma}\right)^{1/(1-\sigma)}, \tag{5}$$

where N^i is the number of goods produced in country i and exported to j.

To obtain total exports from country i to j, we add up over all N^i goods produced in country i and

G

exported to j, obtaining $X^{ij} = N^i p^{ij} c^{ij}$, or from (4) and (5),

$$X^{ij} = N^i Y^j \left(\frac{p^{ij}}{P^j}\right)^{1-\sigma}. \tag{6}$$

Equation (6) is not quite a gravity equation because it does not have the GDP of the exporting country i. There are several methods that can be used to introduce that variable into (6).

One method is to solve for the number of products N^i in the exporting country, by using a free entry condition from the monopolistic competition model. In the simplest monopolistic competition model, with one factor of production and CES preferences, the number of products will be proportional to GDP, so that $N^i \propto Y^i$. We can also introduce transport costs into (6) by writing the c.i.f. prices as:

$$p^{ij} = T^{ij} p^i, \tag{7}$$

where p^i is the f.o.b. (free on board) price from country i and $T^{ij} \geq 1$ is the transport costs from country i to j. By substituting these relations into (6), and taking natural logs, we obtain:

$$\ln\left(X^{ij}\right) = a + \ln\left(Y^i Y^j\right) + (1-\sigma)\ln T^{ij} + (1-\sigma)\ln\left(p^i/P^j\right). \tag{8}$$

Baier and Bergstrand (2001) estimate a gravity equation like (8), using price data for the term $\ln(p^i/P^j)$, and splitting the term $\ln T^{ij}$ into tariffs and transport costs. They use data for Organisation for Economic Co-operation and Development (OECD) countries, and take differences between the averages in 1958–60 and 1986–88 to express all the variables as changes over time. The resulting linear regression has an R^2 of 0.40, so they are able to explain nearly one-half of the growth in exports between OECD countries.

A second method to convert (6) into a gravity equation is to solve for the f.o.b. prices p^i in each exporting country. Anderson and van Wincoop (2003) argue that an implicit solution for the f.o.b. prices is:

$$\tilde{p}^i \equiv \left(s^i/N^i\right)^{1/(1-\sigma)}/\tilde{P}^i, \tag{9}$$

in which case the price indexes are solved as:

$$\left(\tilde{p}^j\right)^{1-\sigma} = \sum_{i=1}^{C} s^i \left(T^{ij}/\tilde{P}^i\right)^{1-\sigma}. \tag{10}$$

If we substitute (9) into (6), we obtain the gravity equation:

$$X^{ij} = s^i Y^j \left(\frac{T^{ij}}{\tilde{P}^i \tilde{P}^j}\right)^{1-\sigma} = \left(\frac{Y^i Y^j}{Y^w}\right)\left(\frac{T^{ij}}{\tilde{P}^i \tilde{P}^j}\right)^{1-\sigma}. \tag{11}$$

So we obtain a gravity equation like (1), where the transport costs T^{ij} can depend on distance d^{ij}. The 'constant' A in the gravity Eq. (11) is $A = \left(\tilde{P}^i \tilde{P}^j\right)^{1-\sigma}/Y^w$, which depends on the price indexes of the exporting and importing country. So the important lesson from Anderson and van Wincoop is that the 'constant' term must vary with the importer and exporter (and if multiple years are used in the data, then it should also vary with time).

Anderson and van Wincoop apply their version of the gravity equation to Canada–US trade. McCallum (1995) originally estimated a gravity equation for trade between ten Canadian provinces, and trade between those provinces and 30 US states, using 1988 data. An updated version using 1993 data, which also includes trade between US states, is as follows (with standard errors in parentheses):

$$\ln X^{ij} = \underset{(0.11)}{2.75}\ \text{Canada} + \underset{(0.05)}{0.40}\ \text{US} + \underset{(0.02)}{1.13}\ \ln\left(Y^i\right) + \underset{(0.02)}{0.97}\ \ln\left(Y^j\right) - \underset{(0.03)}{1.11}\ \ln\left(d^{ij}\right), \tag{12}$$

$$R^2 = 0.85, \quad N = 1511.$$

The first variable appearing on the right of (12), 'Canada', is an indicator variable equal to 1 for trade between Canadian provinces, and zero otherwise. The second variable, 'US', is an

indicator variable equal to 1 for trade between US states, and zero otherwise. The remaining variables are GDPs of the exporting province or state, both of which have coefficients close to unity, and distance, with a coefficient close to minus 1.

Since the variables in (12) are in natural logs, we can take the exponential of the indicator coefficients to obtain $e^{2.75} = 16$ and $e^{0.4} = 1.5$. Thus, the estimates imply that cross-provincial trade within Canada is 16 times greater than cross-border trade with the United States, whereas cross-state trade within the United States is only 1.5 times greater than cross-border trade. (Using data from 1988, McCallum had found that cross-provincial trade within Canada was 22 times greater than cross-border trade.) The very large magnitude of the 'border effect', leading to much more trade *within* Canada than across the border, is very surprising!

Before we try to explain where this border effect might come from, we should check that the estimates in (12) are reliable. In particular, this estimate does not incorporate the price indexes $\tilde{P}^i$ and $\tilde{P}^j$ that appear in (11). If we incorporate these terms, the estimate for 1993 from Anderson and van Wincoop (2003) is:

$$\begin{aligned} \ln\left(X^{ij}/Y^iY^j\right) &= -\underset{(0.08)}{1.65}\ \text{Border} - \underset{(0.03)}{0.79}\ \ln\left(d^{ij}\right) \\ &\quad + \ln\left(\tilde{P}^i\right)^{\sigma-1} + \ln\left(\tilde{P}^j\right)^{\sigma-1}, \\ N &= 1511. \end{aligned} \tag{13}$$

Notice that Anderson and van Wincoop (2003) keep the GDPs on the left of (13), by dividing both sides of (11) by (Y^iY^j). On the right, they include an indicator variable 'Border' equal to 1 for cross-border trade, and zero otherwise, rather than the separate Canada and US indicator variables. In addition, they include the price indexes $\left(\tilde{P}^i\right)^{\sigma-1}$ and $\left(\tilde{P}^j\right)^{\sigma-1}$ on the right. Those price indexes are computed from the formula in (10), where the transport costs used in that formula are:

$$\begin{aligned} (1-\sigma)\ln T^{ij} &= -1.65\ \text{Border} \\ &\quad - 0.79\ln\left(d^{ij}\right). \end{aligned} \tag{14}$$

Thus, we use the estimated coefficients of the border and distance from (13), substitute these into (14) and (10) to compute the price indexes $\left(\tilde{P}^i\right)^{\sigma-1}$ and $\left(\tilde{P}^j\right)^{\sigma-1}$, use these price indexes in the regression (13) to get a new coefficients of the border effect and distance, use those again in (14) and (10) to get new price indexes, and iterate on this procedure until it converges.

The estimated 'border effect' from (13) is $e^{1.65} = 5.2$. So, on average, the Canada–US border leads to five times more trade within the United States and Canada than cross-border trade. This procedure does not directly give an estimate of the *separate* Canada and US border effects, but that can be obtained from an extra calculation using (10) and computing what trade within each country would be with and without the border effect. Anderson and van Wincoop (2003) conclude that trade *within* Canada is 10.5 times higher than cross-border trade due to the border effect, which is smaller than the estimate of 16 times obtained from regression (12) (or the original estimate of 22 times from McCallum 1995). In addition, trade *within* the United States is 2.6 times higher than cross-border trade due to the border effect.

This procedure due to Anderson and van Wincoop requires custom programming to compute the price indexes from (10). Feenstra (2004, ch. 5) notes that a linear regression can be used instead, by estimating the price indexes in (13) using fixed effects, that is, using indicator variables for each exporting region and each importing region whose coefficients are the *estimated* price indexes $\left(\tilde{P}^i\right)^{\sigma-1}$ and $\left(\tilde{P}^j\right)^{\sigma-1}$. That approach gives results very similar to (13), and is easier to compute, so using fixed effects for importing and exporting regions can be considered the preferred estimation method.

The use of fixed effects in the gravity equation has now become common practice, and makes a difference. For example, Rose (2000) estimates a gravity equation across a broad sample of

countries, some of which belong to a currency union. The indicator variables for members of the currency union take on a surprisingly large value, implying that a currency union increases trade between its members by three times, or 200 per cent! That result seems too large. Rose and van Wincoop (2001) include fixed effects for exporting and importing countries, and subsequent papers have also included the fixed effects while allowing them to vary over time (as needed when multiple years are included in the estimation). Baldwin (2006a, b) surveys these papers, and concludes that a currency union increases trade by nearly two times, or 100 per cent. But the impact of the euro on trade is much lower – in the range of eight to 15 per cent.

The use of fixed effects does not necessarily lower the indicator variable coefficients. In another application, Rose (2004) estimates the impact of World Trade Organization (WTO) membership on trade between countries, and finds that it is surprisingly small. He added together exports and imports, using $\ln(X^{ij} + X^{ji})$ as the dependent variable (which is appropriate only in the simplest gravity model without price effects, as in (3), and using a single set of country fixed effects. But Subramanian and Wei (2003) argue that, when the $\ln X^{ij}$ is used as the dependent variable and both exporter and importer fixed effects are included, the WTO has a substantial effect on imports, especially for the industrial countries.

Along with included fixed effects in the estimation (and allowing them to vary over time), another very important estimation issue is how 'zero's' are treated in the data. Notice that we use the natural log of exports as the dependent variable in the gravity equation, so when exports are zero the natural log cannot be computed. Common practice has been to omit those observations, as was done in (12) and (13). In order to incorporate the zero trade flows, Silva and Tenreyro (2006) recommend that estimation be performed as if the dependent variable had a Poisson probability distribution. The Poisson distribution is ordinarily applied to 'count' data, that is, the non-negative integers 0, 1, 2, 3, ... But it can still be used as the estimation method for a continuous, non-negative dependent variable like X^{ij}. In that case the zero trade values can be included.

An alternative approach to incorporate zero trade values is advocated by Helpman, Melitz and Rubenstein (2007). They first model the *reason* why countries might not trade with each other, that is, because of fixed costs of exporting to another country. That model leads to a selection equation for whether countries trade or not, followed by a gravity equation when trade is positive. This two-equation system is estimated using a modified Heckman procedure (that is, one equation for whether a country has positive trade or not, and a second equation explaining the amount of trade). They argue that the estimates from the two-equation system are quite different from a single equation approach that ignores the zero trade flows.

We conclude by noting that the gravity equation is more general than we have indicated so far, in several respects. First, it can be derived even for trade in homogeneous goods (so that countries produce the same good). Eaton and Kortum (2002) derive a gravity equation in a general model with homogeneous goods and many countries, while Evenett and Keller (2002) and Feenstra et al. (2001) also obtain this equation in more specialized settings. Second, the gravity equation can also be used with dependent variables other than trade. For example, Eaton and Tamura (1994) and Head and Ries (2005) derive and estimate a gravity equation that uses foreign direct investment as the dependent variable. Engel and Rogers (1996) estimate a gravity equation that use the variance in prices across cities as the dependent variable, and find that crossing the Canada–US border adds as much to the volatility of prices as adding 2,500 miles between cities. These examples illustrate the many applications of the gravity equation, which will continue to be a widely used empirical tool in international economics.

See Also

- ▶ Border Effects
- ▶ Currency Unions
- ▶ International Trade Theory

Acknowledgment The author would like to thank Keith Head for his very helpful comments.

Bibliography

Anderson, J.A. 1979. A theoretical foundation for the gravity equation. *American Economic Review* 69: 106–116.

Anderson, J.A., and E. van Wincoop. 2003. Gravity with gravitas: A solution to the border puzzle. *American Economic Review* 93: 170–192.

Baier, S., and J.H. Bergstrand. 2001. The growth of world trade: Tariffs, transport costs, and income similarity. *Journal of International Economics* 53: 1–27.

Baldwin, R. 2006a. The euro's trade effects. Working Paper No. 594, European Central Bank.

Baldwin, R. 2006b. *In or out: Does it make a difference? an evidence based analysis of the trade effects of the euro*. London: CEPR.

Bergstrand, J.H. 1985. The generalized gravity equation, monopolistic competition, and the factor-proportions theory in international trade. *The Review of Economics and Statistics* 71: 143–153.

Bergstrand, J.H. 1989. The gravity equation in international trade: Some microeconomic foundations and empirical evidence. *The Review of Economics and Statistics* 67: 474–481.

Bhagwati, J.N. 1972. The Heckscher–Ohlin theorem in the multi-commodity case. *Journal of Political Economy* 80: 1052–1055.

Davis, D.R. 1995. Intra-industry trade: A Heckscher–Ohlin–Ricardo approach. *Journal of International Economics* 39: 201–226.

Eaton, J., and S. Kortum. 2002. Technology, geography and trade. *Econometrica* 70: 1741–1780.

Eaton, J., and A. Tamura. 1994. Bilateral and regionalism in Japanese and U.S. trade and direct foreign investment patterns. *Journal of the Japanese and International Economies* 8: 478–510.

Engel, C., and J.H. Rogers. 1996. How wide is the border? *American Economic Review* 86: 1112–1125.

Evenett, S., and W. Keller. 2002. On theories explaining the success of the gravity equation. *Journal of Political Economy* 110: 281–315.

Feenstra, R. 2004. *Advanced International Trade*. Princeton: Princeton University Press.

Feenstra, R.C., J.R. Markusen, and A.K. Rose. 2001. Using the gravity equation to differentiate among alternative theories of trade. *Canadian Journal of Economics* 34: 430–447.

Head, K., and J. Ries. 2005. Judging Japan's FDI: The verdict from a dartboard model. *Journal of the Japanese and International Economies* 19: 215–232.

Helpman, E. 1987. Imperfect competition and international trade: Evidence from fourteen industrial countries. *Journal of the Japanese and International Economies* 1: 62–81.

Helpman, E., M.Melitz, and Y.Rubinstein. 2007. Trading partners and trading volumes. Working Paper No. 12927. Cambridge, MA: NBER.

McCallum, J. 1995. National borders matter. *American Economic Review* 85: 615–623.

Rose, A.K. 2000. One money, one market: Estimating the effect of common currencies on trade. *Economic Policy* 30: 9–45.

Rose, A.K. 2004. Do we really know that the WTO increases trade? *American Economic Review* 94: 98–114.

Rose, A.K., and E. van Wincoop. 2001. National money as a barrier to international trade: The real case for currency union. *American Economic Review* 91: 386–390.

Silva, J.S., and S. Tenreyro. 2006. The log of gravity. *The Review of Economics and Statistics* 88(4): 641–658.

Subramanian, A. and S.-J.Wei. 2003. The WTO promotes trade, strongly but unevenly. Working Paper No. 10024. Cambridge, MA: NBER.

Tinbergen, J. 1962. *Shaping the world economy*. New York: Twentieth Century Fund.

Gravity Models

Pierre-Philippe Combes

Abstract

Basic gravity models state that economic interactions between two geographically defined entities are proportional to the size of these entities and inversely related to the distance between them. They have great empirical explanatory power. The impact of distance is strong and not diminishing over time. Extended gravity models incorporate borders and contiguity effects and more sophisticated interaction cost measures. They can be theory grounded, which makes each country's location vis-à- vis the rest of the world play a role in the bilateral relationship. Various empirical approaches have been proposed to tackle the econometric issues at stake in these more sophisticated frameworks.

Keywords

Borders; Business and social networks; Comparative advantage; Distance; Equity flows; Fixed-effects; Foreign direct investment; Gravity models; Heteroskedasticity bias; Imperfect competition; Information costs;

Market potential; Maximum likelihood; Microfoundations; Migration; Monopolistic competition; Nonlinear estimation; Patents; Preference bias; Proximity; Trade agreements; Reverse causality; Selection bias; Social interaction; Social networks; Spatial interaction; Trade; Trade costs; Transport costs

JEL Classifications
F1; F2; R12; R23; R40

The simplest gravity model states that economic or social interactions between two geographically defined economic entities are proportional to the size of these entities and inversely related to the distance between them. The system of interactions that results from these bilateral relationships shapes the spatial organization of the global economy. Initially borrowed from the universal law of gravitation that Newton established in 1687 for heavenly bodies, this model has undergone many refinements in economics, in particular in order to better match some underlying theoretical models and data. Some of the first applications were proposed by social physicists such as Ravenstein (1885) for migration flows, and Reilly (1931) for consumers' shopping behaviour. Stewart (1947), an astronomer, suggested that the gravity law could be applied to a very wide class of social interactions. Tinbergen (1962) initiated what continued to be the main application of gravity models, namely, the study of the determinants of trade.

The Basic Framework

Typically, country i's exports to country j, F_{ij}, are modelled as a function of the distance between these countries, d_{ij}, and of their economic mass, (M_i, M_j), which is most often proxied by their GDPs. Thus a basic gravity trade model estimates parameters α, β, and δ (expected to be positive) such that

$$F_{ij} = G\frac{M_i^{\alpha}M_j^{\beta}}{d_{ij}^{\delta}}\varepsilon_{ij},$$

where G is a constant and ε_{ij} an error term capturing what is left unexplained by the model. M_i can be interpreted as the supply of the good traded and M_j the demand, while d_{ij}^{δ} captures trade costs, which encompass all costs incurred in transferring goods. These costs add to the price of goods when they are not sold locally and are assumed to increase with spatial distance. Alternatively, if F_{ij} is the number of migrants from i to j, regional populations are often more relevant as measures of M_i and M_j, while d_{ij}^{δ} reflects a moving cost. Similar interpretations can be proposed for other kinds of flows.

According to gravity models, proximity is the main engine of trade, of migration or of any precisely defined social interaction between spatially distinct economic entities. This could appear as an obsolete view of the world if one believes in the 'death of distance', as touted by popular accounts. However, Disdier and Head's (2008) meta-analysis over 1,467 estimations of δ on trade flows indicates an average value around 0.9: halving distance increases trade by 45 per cent. These authors report even larger δ's for recent periods, which means that the distance decay effect has actually increased in recent years.

In addition to the reliable estimates of the impact of distance they lead to, the success of gravity models is due to their great explanatory power for flows, and this holds true whatever the geographical scale (countries, large or small regions), the period of study or the goods considered. This makes gravity one of the most stable relationships in economics and a useful predictive tool. It can be also used for obtaining predictors of variables used in a second stage to explain income, productivity, or growth dispersion across space (for example, Redding and Venables 2004).

A Wide Range of Applications

Gravity models are applicable to many other endogenous quantities in addition to trade flows. We have mentioned migration. Urban planners use them for traffic forecasting. There is also evidence of gravity effects in explaining foreign

direct investment. More novel are the estimations on equity flows. In that case, the explanatory power of the model is as great as it is for trade, and halving distance increases flows by 25 per cent (Portes and Rey 2005). In a supposedly financially integrated world, this is high. The impact of distance is still significant, although around six times lower, for flows of ideas, identified for instance as the citations of patents in a country that have been taken out in another country (Peri 2005).

McCallum (1995) shows that borders also matter a lot, as well as distance. Trade between two Canadian regions was found to be 20 times larger than trade between a Canadian region and a US state of the same size and at the same distance, even if this drops to around six times once some statistical problems are removed. Discrete gaps in trade flows are also systematically observed between areas that are contiguous, relative to those that are not. Such effects suggest that the impact of distance on trade is not log-linear, or even smooth, and that a wider class of spatial proximity measures is necessary to fully encompass the effects of space.

Proximity matters for spatial interactions because it proxies for many of their determinants. Transport costs are the most obvious one. Clearly, the energy consumption or the time spent in transport, which results in opportunity costs, increase with distance. For migration, moving costs (both monetary and psychological) increase with distance and jump upward once borders are crossed. International trade flows are clearly reduced by trade policies, but trade agreements are typically first established between nearby countries. More original is the idea that preferences and tastes may be biased towards local goods, which may result from better information about them. Information costs are also critical for firms that want to access distant markets. They need to find local retailers, and then to work with them (which also clearly matters for foreign direct investment). Prior to this, they must assess market size, find out about local tastes and possibly adapt their products and marketing. All of this may explain not only the impact of distance but also the role of additional proximity measures. Moreover, on top of a possible composition effect (relatively more goods that are less easy to trade are traded), this suggests another explanation for the recently increasing impact of distance. While transport costs and trade barriers have clearly strongly diminished, preference biases or information costs may have risen, possibly due the increasing number and complexity of the goods available.

Simple gravity models have been extended to control for economic factors that would directly capture trade costs. Large data-sets on trade agreements or on transport costs are now available. The fraction of the population sharing the same language is used to capture closeness of tastes or reduced information costs. More generally, the positive role of business and social networks (among migrants from the same country or among firms belonging to the same business groups) on trade is currently being studied. These measures reduce the impact of distance, borders, and contiguity, without completely eliminating them. Naturally, not all the reasons why space matters for interactions have yet been identified. Extensive references on trade costs proxies and the various applications of gravity models can be found in Anderson and van Wincoop (2004) and in Combes et al. (2008, ch. 5).

Extended Frameworks and the Role of the Rest of the World

The bilateral nature of the gravity model is somewhat surprising. One would expect the actual location of the respective economies vis-à-vis the rest of the world to matter also. Trade between Australia and New Zealand is certainly much larger than it would be if these countries were not so isolated. But nothing in basic gravity models takes this into account. Similarly, if the relationship results from equilibrium between supply and demand, the absence of the goods prices in the model is also striking. Moreover, economists have long been challenged by the strong empirical support of gravity models and, simultaneously, their lack of theoretical foundations. Therefore, from Anderson (1979) on, a number of approaches have been proposed to

derive the gravity relationship from fully specified theoretical models. Anderson and van Wincoop (2004) show that for trade flows the class of models that can be used is fairly large. They are based on either comparative advantage or on imperfect competition; the most popular uses monopolistic competition. They all lead to specifications more complex than (1) but such microfoundations significantly improve the understanding of gravity models and of their underlying mechanisms. Their main feature is that the interactions between two economic entities, such as regions, do depend on their location relative to other areas because of interesting price effects. Indeed, in an economy with more than two regions and costly trade, the supply, demand and price of goods traded between two given economies depend on the relative costs of all firms and consumers to access all the markets. For instance, if Australia and New Zealand reduced their trade costs to all other developed countries, the price of the goods they exchange bilaterally would increase relative to the price of all other goods, which would reduce their bilateral trade. On the other hand, the overall saving in trade costs and the increase in competition it induces would also imply a greater purchasing power, which would increase trade with all partners, including bilateral trade. Interestingly, in these models local incomes can be shown to be function of the area's market potential. This potential takes a form similar to the economic version proposed by Harris (1954), which is reminiscent of gravitational or electric potentials, which must be enriched by price effects again. This is the sum of all regions' income discounted by trade costs and weighted by complex price effects.

Econometric Issues

Unfortunately, such formulations modify (1) in ways that make estimation more cumbersome. Typically the equation becomes nonlinear in some unknown parameters (such as price elasticity) that must be estimated simultaneously with the impact of trade costs. Various approaches, detailed in Feenstra (2004, ch. 5), have been proposed to deal with that difficulty. The first one consists of using nonlinear estimation procedures. Another solution takes as the left-hand-side variable the ratio of the bilateral flow to the flow to a reference destination, which makes the right-hand-side variables depend again on the origin and destination only. People sometimes use real price data, but these rarely match their exact definitions in the theoretical model. The least data-demanding strategy, which remains compatible with a large number of theory-grounded approaches, consists in estimating fixed effects for each origin and destination. However, the impact of the determinants of the fixed effects, the mass of the economies and their locations for instance are no longer identified.

Other econometric problems remain. First, the fact that each country trades with many destinations induces correlations between error terms, and therefore possible heteroskedasticity biases. More problematic are the zero trade flows towards a large number of destinations that often characterize many countries, which may induce selection biases. Santos Silva and Tenreyro (2006) propose a pseudo-maximum likelihood estimator to deal with both issues. The literature based on heterogeneous firms and the presence of fixed export costs might help provide more theory- grounded approaches to these problems (Helpman et al. 2008). The fact that economic masses and prices are simultaneously determined with trade flows might also bias the estimations. Appropriate instrumental procedures should help, however. Last, the trade cost proxies might themselves be endogenous. For instance, new infrastructure can be built in response to an increase of trade flows, trade agreements may be signed preferably between privileged trade partners and networks may emerge once trade is large. This is more difficult to handle and these possible reverse-causality biases have yet to be really investigated.

Hence, the long history of gravity models does not prevent them from stimulating a lot of current research, both theoretical and empirical.

Microfounded frameworks move the model further and further away from Newton's law, adding a lot to the understanding of the mechanisms shaping spatial interactions. Additional challenges lie ahead.

See Also

- ▶ GIS Data in Economics
- ▶ International Trade and Heterogeneous Firms
- ▶ New Economic Geography
- ▶ Regional Development, Geography of
- ▶ Spatial Economics
- ▶ Systems of Cities
- ▶ Tradable and Non-Tradable Commodities
- ▶ Trade Costs

Bibliography

Anderson, J. 1979. A theoretical foundation for the gravity equation. *American Economic Review* 69: 106–116.

Anderson, J., and E. van Wincoop. 2004. Trade costs. *Journal of Economic Literature* 42: 691–751.

Combes, P.-P., T. Mayer, and J.-T. Thisse. 2008. *Economic geography*. Princeton: Princeton University Press.

Disdier, A.C., and K. Head. 2008. The puzzling persistence of the distance effect on bilateral trade. *The Review of Economics and Statistics* 90: 37–48.

Feenstra, R.C. 2004. *Advanced international trade*. Princeton: Princeton University Press.

Harris, C. 1954. The market as a factor in the localization of industry in the United States. *Annals of the Association of American Geographers* 64: 315–348.

Helpman, E., M. Melitz, and Y. Rubinstein. 2008. Estimating trade flows: Trading partners and trading volumes. *Quarterly Journal of Economics* 123: 441–487.

McCallum, J. 1995. National borders matter: Canada-US regional trade patterns. *American Economic Review* 85: 615–623.

Peri, G. 2005. Determinants of knowledge flows and their effects on innovation. *The Review of Economics and Statistics* 87: 308–322.

Portes, R., and H. Rey. 2005. The determinants of cross-border equity flows. *Journal of International Economics* 65: 269–296.

Ravenstein, E.G. 1885. The laws of migration. *Journal of the Royal Statistical Society* 48: 167–227.

Redding, S., and A. Venables. 2004. Economic geography and international inequality. *Journal of International Economics* 62: 53–82.

Reilly, W.J. 1931. *The law of retail gravitation*. New York: Pilsbury.

Santos Silva, J., and S. Tenreyro. 2006. The log of gravity. *The Review of Economics and Statistics* 88: 641–658.

Stewart, J.Q. 1947. Suggested principles of 'social physics'. *Science* 106: 179–180.

Tinbergen, J. 1962. *Shaping the world economy: Suggestions for an international economic policy*. New York: Twentieth Century Fund.

Gray, Alexander (1882–1968)

T. Johnston

Gray was a civil servant (1905–21), professor of political economy (Aberdeen University, 1921–35; University of Edinburgh, 1935–56) and, throughout his career, a public servant, poet and translator. Gray combined an early training in mathematics with Scottish pragmatism and an international outlook developed as a graduate student at Göttingen and Paris. He was not interested in esoteric economic theory, but in applications, in public policy and in the historical development of ideas. His early writings (1923, 1927) drew on his civil service experience in establishing the welfare state. Later, in his lectures, he became a leading analyst of the growth of the nationalized industries in Britain. His writings in the history of economic thought (1931, 1946) were remarkable for their range, dissecting ideas from Ancient Greece to modern times and showing a profound knowledge of the literature in many languages. He did not espouse any particular thesis about the development of ideas, but was temperamentally critical of socialism and of the growth of state intervention. He was an inspiring teacher, setting economics in context, historically and politically. Outside economics his translations into broad Scots of European ballads and of Heine were sensitive and much admired.

Selected Works

1923. *Some aspects of national health insurance*. London: P.S. King & Son.

1929. *Family endowment*. London: Ernest Benn.
1931. *The development of economic doctrine*. London: Longmans, Green & Co.
1946. *The socialist tradition. Moses to Lenin*. London: Longmans, Green & Co.

Gray, John (1799–1883)

N. W. Thompson

From the time that he went to work in a London manufacturing and wholesale house at the age of 14, Gray was, apparently, interested in social reform and, after attending the London tavern debates of 1817, his thinking began to assume an Owenite socialist complexion. It was this interest in cooperative socialism that led him to visit the community established by Abram Combe at Orbiston in 1825, though his reaction as expressed in *A Word of Advice to the Orbistonians* (1826) was critical. Nevertheless his first major work, *A Lecture on Human Happiness* (1825), did embrace the communitarian ideal. In it he argued that under competitive capitalism the real income of the country, which consisted of the quantity of wealth annually produced by the labour of the people, was taken from its producers through the rent, interest and profits of those who bought labour at one price and sold it at another. This failure to exchange equivalents would be eliminated through the formation of cooperative communities and the abolition of the competitive system of exchange itself.

Yet if it was the *Lecture* which secured him a measure of contemporary notoriety, *The Social System: A Treatise on the Principle of Exchange* (1831) is Gray's most interesting work. Here his explanation of the impoverishment of labour is similar to that of the earlier book. Exploitation was deemed to occur in the sphere of exchange, producing poverty, distress, underconsumption and thence arrested economic development. Gray's solution, however, represented a significant move away from his earlier communitarianism in the direction of a centrally controlled, technocratically run economy, more Saint-Simonian than Owenite in character.

As in the *Lecture*, the competitive market economy was to be abolished but in *The Social System* its pricing, allocative, distributive and equilibrating functions were to be performed by a central authority denominated by Gray the National Chamber of Commerce. Thus

> the Social System recognises as useful, but one controlling and directing power, but one judge of what is prudent and proper to bring into the market, either as respects kind or quantity – the Chamber of Commerce – who, having the means of ascertaining, at all times, the actual stock of any kinds of goods in hand would always be able to say at once where production should proceed more rapidly, where at its usual pace and where also it should be retarded.

Further the Chamber would employ 'agents' to organize the 'cultivation of the land and the management of all those trades and manufactures which had been brought together by a voluntary association of landowners, capitalists and traders'. It would also be responsible for ensuring that demand was commensurate with supply, for making, as Gray phrased it, 'Production the uniform and never failing cause of demand.'

The central themes of *The Social System* were to be reiterated in a work entitled *An Efficient Remedy for the Distress of Nations* (1842), two works which taken together represent the first significant attempt, however inadequate, in the history of British socialist thought to discuss how central control or authority might be applied to a modern, complex, interdependent, industrial economy.

Yet by 1848, and the publication of his *Lectures on the Nature and Use of Money*, Gray had abandoned both communitarianism and central economic planning, arriving instead at the conclusion that the system of exchange might be rationalized without any transcendence or abolition of the market and without the institutional paraphernalia of communities or chambers of commerce. There was now no problem as there had been in the *Lecture*, in reconciling free and

equitable exchange with free competition. 'The great principle of individual competition should be left free and unfettered as the air we breathe.' All that was required to make competitive capitalism work, all that was needed to ensure generalized prosperity was a medium of exchange that could be expanded *pari passu* with the level of output. This would eliminate the incidence of general economic depressions and lay the basis for rising living standards and social harmony. As Gray saw it 'A few salutary money laws are all that are wanted' because 'our false monetary position is the one and only cause of our misfortunes'.

By his death in 1883, after a successful career in publishing, Gray had become reconciled to market capitalism at a practical as well as a theoretical level.

Selected Works

1825. *A lecture on human happiness*. London.
1826. A Word of Advice to the Orbistonians on the principles which ought to regulate their present proceedings, 29 June 1826. Edinburgh.
1831. *The social system: A treatise on the principle of exchange*. Edinburgh.
1842. *An efficient remedy for the distress of nations*. Edinburgh.
1848. *Lectures on the nature and use of money*. Edinburgh.

References

Beales, H.L. 1933. *The early English socialists*. London: Hamish Hamilton.
Beer, M. 1953. *A history of British socialism*, 2 vols. London: Allen & Unwin.
Cole, G.D.H. 1977. *A history of socialist thought*, 5 vols. Vol. 1: *Socialist thought: The forerunners, 1789–1850*. London: Macmillan.
Foxwell, H.S. 1899. Introduction to the English translation of A. Menger, *The right to the whole produce of labour*. London: Macmillan.
Gray, A. 1967. *The socialist tradition, Moses to Lenin*. London: Longman.
Hunt, E.K. 1980. The relation of the Ricardian socialists to Ricardo and Marx. *Science and Society* 44: 177–198.
Kimball, J. 1946. *The economic doctrines of John Gray 1799–1883*. Washington, DC: Catholic University of America Press.
King, J.E. 1981. Perish Commerce! Free trade and underconsumption in early British radical economics. *Australian Economic Papers* 20: 235–257.
King, J.E. 1983. Utopian or scientific? A reconsideration of the Ricardian socialists. *History of Political Economy* 15: 345–373.
Lowenthal, E. 1911. *The Ricardian socialists*. New York: Longman.
Martin, D. 1982. John Gray 1799–1883. In *Dictionary of labour biography*, vol. 6, ed. J. Saville and J. Bellamy. London: Macmillan.
Thompson, N.W. 1984. *The people's science: The popular political economy of exploitation and crisis, 1816–34*. Cambridge: Cambridge University Press.

Gray, Simon (Alias George Purves, LL.D.) (fl. 1795–1840)

Peter Newman

Little is known about Gray except that he worked at the War Office in the late eighteenth and early nineteenth centuries. His major work (if that is the right phrase) was *The Happiness of States*, a book which 'he meant to have published in 1804 but was prevented', with the result that it did not appear until 1815. A large and cranky work, it is at once anti-Physiocrat, anti-Smith (the similarity of its title to *The Wealth of Nations* is no accident) and anti-Malthus.

Its reception was so poor that in 1817 and 1818, under the pseudonym of George Purves LL.D., he published two other works whose main purpose was to praise his own *Happiness of States* (see Sraffa 1952, p. 38n). Such self-puffery was of course not unknown, even for works of the first magnitude. (When his *Treatise of Human Nature* 'fell "dead-born from the press"' in 1739, the great but dejected Hume responded by publishing anonymously a year later *An Abstract of a Book lately Published*; *Entituled*, *A Treatise of Human Nature*, &*c*., in which he praised his otherwise ignored work

(Keynes and Sraffa 1938).) Whilst Hume's puff was unsuccessful (no second edition of the *Treatise* appearing in his lifetime) Gray's was not, at least in his own judgement, since in 1819 he brought out a second edition of his original work.

Gray could be left in well-deserved obscurity were it not for one thing. Chapter V of Book VII of *Happiness*, entitled 'Of Scarcities', contains a most detailed account of what economists usually call Giffen goods. 'To raise the price of corn in any great degree, tends directly to increase the general consumption of that necessary (p. 505). 'There is no paradox here', since bread is both an inferior good and a large item in the budgets of the poor. Thus 'By raising the price of bread corn,. . . we force them to have more on it; . . . However, paradoxical, therefore, it may be in seeming, it is a plain substantial fact, that the higher the price of corn and potatoes, the greater is the consumption . . .' (pp. 509–10, original punctuation).

Perhaps we should call them not Giffen but Gray goods.

Selected Works

1797 *The essential principles of the wealth of nations, illustrated, in opposition to some false doctrines of Dr. Adam Smith, and others.*. . . London: T. Becket.

1815 *The happiness of states: Or an inquiry concerning population. The modes of subsisting and employing it, and the effects of all on human happiness*. London: Hatchard. 2nd edn., 1819.

1817 *[As George Purves]. All classes productive of national wealth,*. . .. London: Longman. 2dn edn, 1840.

1818 *[As George Purves]. Gray versus Malthus, the principles of population and production investigated,*. . .. London: Longman etc.

1820 *Remarks on the production of wealth . . . in a letter to the Rev. T.R. Malthus,*. . .. London.

1823 *[As George Purves] The grazier's ready reckoner, being a complete set of tables, shewing the weight of cattle, calves and pigs by admeasurement, etc*. Warrington.

1839 *The Spaniard; or, Relvindez and Elzora, a tragedy [in five acts and in verse], and the young country widow [in five acts and in prose],*. . .. London.

1842 *The Messiad; or, the life, death, resurrection, and exaltation of Messiah, the Prophet of the Nations*. London.

References

Keynes, J.M., and P. Sraffa (eds.). 1938. *An abstract of a treatise of human nature 1740, by David Hume*. Cambridge: Cambridge University Press.

Masuda, E., and P. Newman. 1981. Gray and Giffen goods. *Economic Journal* 91: 1011–1014.

Powell, E.G. 1896. Gray, Simon. In *Dictionary of political economy*, vol. II, ed. R.H.I. Palgrave, 257–258. London: Macmillan.

Sraffa, P. (ed. with the collaboration of M.H. Dobb). 1952. *The works and correspondence of David Ricardo*, vol. VIII. Cambridge: Cambridge University Press for the Royal Economic Society.

Great Depression

Peter Temin

Abstract

The world depression of the 1930s was the greatest peacetime economic catastrophe in history. There had been hard times before, but never without war, natural disaster or pestilence. The massive and long-lasting unemployment and hardship of the 1930s was a pathology of industrial society, caused by a malfunctioning of the economic system. Adherence to gold-standard policies led to a set of currency crises in 1931 that turned a bad recession into the Great Depression.

Keywords

Banking crises; Beggar-my-neighbour; Currency crises; Deflation; Demand shock; Devaluation; Federal Reserve System; Friedman,

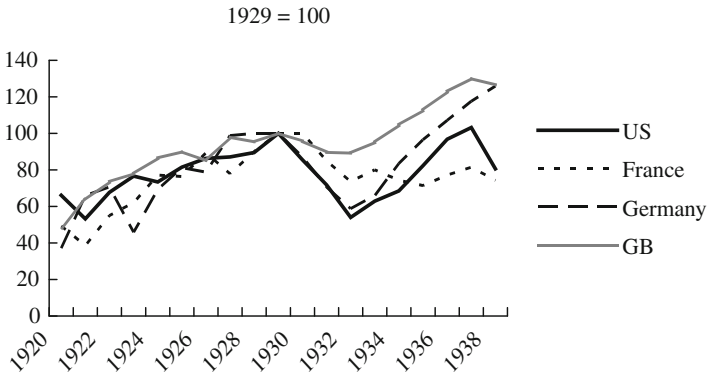
M.; German hyperinflation; Gold standard; Great Depression; Kindleberger, C.; Real wage rates; Schwartz, A.; Specie-flow mechanism; Transfer problem; Unemployment; Young Plan

JEL Classifications
N3

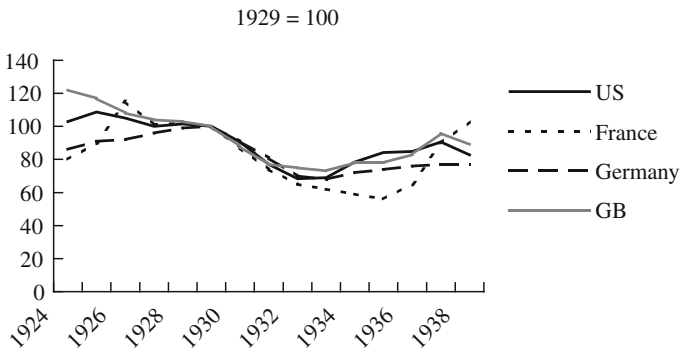
Magnitude

Figure 1 shows the fall in industrial production during the Great Depression in the four largest national economies at that date. Industrial production declined by almost half in the United States and Germany. It fell more slowly and continuously in France, and paused rather than fell in Great Britain. National incomes did not fall as far as industrial production since services did not contract as much, but they decreased sharply; real per-capita GNP in the United States fell by one-third. National experiences in the depression varied greatly, but very few countries in the world escaped the economic hardship of the 1930s. One task for any account of the Great Depression is to explain its worldwide impact.

Figure 2 shows the fall in wholesale prices for the same four countries. Prices fell at the same time as production, by the same amount or more. Unemployment grew dramatically in almost all countries. Rates for the four largest economies are shown in Table 1. Only in the United Kingdom were unemployment rates approximately as high in the 1920s as in the 1930s, due to depressed conditions in Britain during the 1920s and a mild depression in the 1930s. Other countries for which we have data fit the more common pattern of higher unemployment in the 1930s.



Great Depression, Fig. 1 Industrial production, 1920–1939 (Source: Temin (1989))



Great Depression, Fig. 2 Wholesale prices, 1924–1939 (Source: Temin (1989))

Great Depression, Table 1 Industrial unemployment rates, 1921–1938

Country	1921–1929	1930–1938	Ratio
France	3.8	10.2	2.7
Germany	9.2	21.8	2.4
United Kingdom	12	15.4	1.3
United States	7.9	26.1	3.3

Source: Temin (1989)

Unemployment meant distress in the 1930s, most visible in Europe and North America. Diets in Europe became very monotonous despite the presence of homegrown vegetables in some areas. Families ate meat only rarely, starches were the basis of most diets, and sugar frequently was replaced by cheaper saccharine. Even this poor diet consumed almost all the family income. Families with children bought milk, most families bought coal for heat, but there was little money left over for clothes and other expenses. Shoes in particular were a problem. Families typically could not afford to replace shoes that had worn out, and so they were patched and patched again. Some families even restricted the activities of their children to save the wear and tear on their shoes.

While spending was channelled into food, and food into bread and coffee, personal travel was reduced to journeys to local neighbourhoods and villages. Trips to towns and town centres had been increasing during the 1920s, to go to the theatre, do Christmas shopping, or attend school. With unemployment, the money to undertake these journeys vanished. Even tram and train fares became a burden, and people relied more heavily on their bicycles. The isolation of rural villages, alleviated by the railways and the prosperity after the First World War, reappeared in the depression.

Unemployed men were exceedingly idle; an increase of apathy reduced all forms of recreational activity. Men passed their time doing essentially nothing; when asked, they could not even recall what they had done during the day. They sat around the house, went for walks – walking slowly – or played cards and chess. Most men went to bed early; there simply was no reason to stay awake. Women were far more active. They spent time cooking, mending clothes to make them last longer, and managing their budgets. Men contributed less to the running of the household than before, sometimes not even turning up on time for meals, and women had the full responsibility for maintaining the household. Even though women previously had struggled to complete their housework after working, they uniformly would have preferred being back at work.

Sociologists observed that most European unemployed families were resigned to their condition. Such families were hanging on, preserving as much of their life and family as they could on their meagre budgets. All their activity was dedicated to getting by; no thought was given to the future. Some families still planned as before, but others collapsed entirely into mental and physical neglect and conflict.

Beyond Europe and North America, the story of destitution was the same, although the workers' issues typically were more related to physical survival. Rural families in Asia and Africa suffered from the low prices that their crops received in the depressed world markets. They do not seem to have lapsed into idleness like unemployed urban workers, but rather continued to produce crops in the hope of increasing their incomes. Consumers in India, no longer able to afford imported cloth, gave a boost to domestic, beleaguered handloom weavers. Workers in Latin America retreated from cities and organized agriculture back into the countryside, and little is known of their living conditions. Latin American governments divided into active states that tried to insulate their economies from the outside world and passive states that waited for better times. Governments were surprisingly stable under this economic stress, but they collapsed in some countries, ranging from Germany to Burma.

Analysis

The first question to ask about this contraction is whether the shocks that produced it were demand or supply shocks. The simultaneous fall in production and prices indicates that the shocks were demand shocks, that the economies of the world were moving down along their upward-sloping

aggregate supply curves in response to downward shifts of aggregate demand curves. The apathetic reaction to unemployment in the Great Depression confirms the hypothesis that the depression was due to a demand shock. Had it been due to a supply shock, families would have been unemployed by choice, happy with their extra leisure. The psychological depression also put great strains on the social structure, and even the political structure in some countries. It was in soil such as this that the noxious weed of National Socialism grew in Germany.

A second question about the Great Depression is how so many countries could have had negative demand shocks at the same time. The answer is that all these countries were adopting deflationary policies according to the dictates of the gold standard. The gold standard was characterized by the free flow of gold between individuals and countries, the maintenance of fixed values of national currencies in terms of gold and therefore each other, and the absence of an international coordinating or lending organization such as the International Monetary Fund. Under these conditions, the adjustment mechanism for a deficit country was deflation rather than devaluation – that is, a change in domestic prices instead of a change in the exchange rate. Lowering prices and possibly production as well would reduce imports and increase exports, improving the balance of trade and attracting gold or foreign exchange. (This is the price-specie-flow mechanism first outlined by Hume in 1752.)

A recession began at the end of the 1920s in the United States and Germany. Both countries began to contract economically, at least partly as a result of central bank pressure. The initial downturns appear to be independent in each country, but their economies were connected, and it is hard to be sure about this. In any case, it was gold-standard policies that transformed the downturn into the Great Depression and pulled the rest of the world down. The choice of deflation over devaluation was the most important factor determining the depth of the Great Depression. The choice was seen clearly and supported by contemporaries in all industrial countries who insisted that the way out of depression was to cut wages and thereby the costs of production and the prices of goods and services. Devaluation was not a respectable option.

Less developed countries were less likely to be on the gold standard than those in Europe or North America. They suffered from the depression nonetheless because of their ties to gold standard countries. As industrial countries reduced their demand for imports, exports from less developed countries declined. As industrial countries stopped exporting capital, less developed countries found their balance of payments deteriorating further. A few countries, such as Spain and Japan, devalued their currencies early and avoided the worst of the depression, but many more countries were not in a position to do this or where it would have had a large effect.

A third question that economists ask about the Great Depression is why the fall in demand was not absorbed entirely in falling prices. In other words, why did prices not fall more and production less than shown in Figs. 1 and 2? The relative stability of wages caused production and employment to fall; falling prices and wages did not absorb the full brunt of the fall in demand. Falling prices also put pressure on financial institutions, whose failures reduced production as well.

Governments and central banks could not easily deflate their economies in the aftermath of the First World War. Workers, who had borne the burdens of international stability mutely in the past, expected and even demanded a voice in policy after their sacrifices during that war. The inability of economic policymakers to force wages down rapidly created the conditions for the Great Depression. The political strains generated by attempts to lower wages caused investors to fear for the stability of the gold standard even as policymakers struggled to maintain it. One reason the gold standard worked well before 1914 was that labour had no voice. The spread of democracy both cast doubt on the monetary authorities' commitment to the gold standard and reduced price flexibility.

Banks failed right and left in the midst of deflation and currency crises. Widespread banking failures were restricted to countries on the

gold standard, showing that the strain of the gold standard was the principal cause of bank distress. All banks suffered as economic activity and prices declined, but the diversion of central banks from the support of commercial banks to the defence of the currency made the difference between banks in difficulty and banking crises. The German government took over the country's great banks in June and July 1931; American banks were allowed to fail continuously as economic decline continued. It seems that a slow crisis was more destructive of economic activity than a rapid one, though there are not enough observations to test this hypothesis.

Narrative

The narrative of the Great Depression properly begins with the First World War. The dislocations of the war and the peace agreements meant that many adjustments had to be made in the international economy. Strains were evident in the immediate aftermath of the war, resulting in hyperinflations in several countries, most notably Germany. The response was to return to the gold standard in the mid-1920s in the hopes of regaining pre-war stability. Alas, the cure proved worse than the disease.

Federal Reserve policy became contractionary at the start of 1928 in order to combat speculation in the New York stock market and to arrest a gold outflow begun in part by previous financial ease. The gold outflow was a prominent determinant of the policy change, even though it was tiny relative to US reserves. The Federal Reserve's primary aim in 1928 and 1929 was to curb speculation on the stock exchange while not depressing the economy. Even though this policy did not impede stock-market speculation, it reduced the rate of growth of monetary aggregates and caused the price level to turn down. The monetary stringency was even tighter than it seems from examining the aggregate stock of money because the demand for money to effect stock-market transactions rose, leaving less for other activities.

The German economy was heavily dependent on imported capital in the 1920s. Popular history regards the capital imports as a necessary offset to Germany's outflow of war reparations payments; they were needed to solve the transfer problem. The reality was quite different. Germany managed to avoid paying reparations by a variety of economic and political manoeuvres that succeeded in postponing its obligations until they could be repudiated entirely. The capital inflow therefore represented a net increase in the resources available to the German economy. The Reichsbank paradoxically worried that this capital inflow was unhealthy and acted to curtail it, sharply reducing the amount of credit available on the German market at the end of the 1920s. The capital flow from the United States to Germany ceased at the end of the 1920s, but the downturn in Germany preceded this fall and derived largely from German economic policies.

At its inception, the Great Depression was transmitted internationally by a gold-standard ideology, a mentality that decreed that external economic relations were primary and that speculation like the booming stock markets in New York and Berlin was dangerous. As the American, British and German economies contracted, they depressed other economies through the mechanism of the gold standard. These countries reduced their imports as they contracted, reducing exports from other countries. They also reduced their capital exports or increased their capital imports in response to the tight credit conditions at the end of the 1920s.

A bad recession turned into the Great Depression in the summer and autumn of 1931. A series of currency crises led both to what we now regard as perverse policy responses and to failures of financial institutions. A warning came in May 1931 when the main bank of Austria, the Credit Anstalt, failed, taking the Austrian schilling with it. This was a preview of things to come, but not a cause of them. The German mark had been under pressure since the German recession began in the late 1920s and the Weimar government began to run increasingly large deficits. They were covered by foreign lending, of which the American Young Plan was the most famous. The Weimar government, however, scared its foreign creditors by a series of statements for domestic consumption

about a customs union with Austria and a possible repudiation of First World War reparations. The Reichsbank lost reserves precipitously in late May, and free trading in the mark was suspended in July 1931.

The British government found itself in similar trouble as its deficits followed Germany's. The Bank of England, unwilling to raise the bank rate above six per cent and further depress the domestic economy, abandoned the gold standard, floated the pound, and devalued in September 1931. The Federal Reserve, facing similar problems and adverse speculation, chose to raise its discount rate by 200 basis points in October 1931. This dramatic measure saved the dollar but killed the domestic economy. It was, however, loudly applauded by the American financial community as the correct gold-standard action.

The effects of fixed exchange rates can be seen in a comparison of Figs. 1 and 2. Figure 1 shows that industrial production in four major countries declined at quite different rates. Figure 2 shows that the rate of decline in prices in the same four countries was strikingly similar. The fixed exchange rates of the gold standard led to uniform changes in prices even though other factors affected the change in production. The standard deviation of price changes was smaller than the standard deviation of production changes for 21 countries on the gold standard in 1930–2, as shown in Table 2. The standard deviation of price changes was smaller than the standard deviation of changes in the industrial production index in each year, even though the standard deviation of both series rose in 1932 as some countries abandoned gold. The final row of Table 2 shows the standard deviations in 1932 for seven countries that stayed on gold in 1932. Even though data for these countries are indistinguishable from the rest of the sample in 1930 and 1931, they are far more uniform in 1932.

Great Depression, Table 2 Standard deviation of changes in 21 gold-standard countries, 1930–1932

Year	Prices	Industrial production
1930	0.037	0.081
1931	0.055	0.078
1932	0.090	0.123
1932[a]	0.035	0.039

[a]Seven countries still on gold in 1932
Sources: Bernanke and James (1991); Temin (1993)

No country on the gold standard, however large, could escape the discipline of this harsh regime in the depression. In almost all cases, deflation was accompanied by depression as declining aggregate demand moved countries down upward-sloping aggregate supply curves. Banking systems in many gold-standard countries collapsed under this deflationary pressure, further reducing economic activity. The Federal Reserve sharply raised the US discount rate in October 1931 in response to a threatened outflow of gold, even though the US economy was contracting rapidly and had massive gold reserves. The primary transmission channel of the Great Depression was the gold standard.

It follows that abandoning the gold standard was the only way to arrest the economic decline. Going off gold severed the connection between the balance of payments and the domestic price level. Countries could lower interest rates or expand production without precipitating a currency crisis. Changes in the exchange rate rather than changes in domestic prices could eliminate differences between the level of domestic and foreign demand without a painful deflation. Any single devaluation could beggar neighbours under some conditions, but universal devaluation would have increased the value of world gold reserves and allowed worldwide economic expansion.

Great Britain abandoned the gold standard in September 1931 after a speculative attack on the pound prompted by bad budgetary news and by contagion from the German currency crisis of July 1931. Great Britain and the countries that followed Britain off gold were not large enough for their actions to arrest the world decline, and they were criticized at the time for abandoning gold; but the world would have been far better off if others had followed them off gold.

Even in the United States, with its vast economic resources and gold reserves, going off gold was a necessary prerequisite for economic expansion. Great Britain avoided the worst of the Great Depression by going off gold in 1931, as shown in Fig. 1. Spain avoided the depression by never

being on the gold standard; Japan by a massive devaluation in 1932. At the other extreme, the members of the gold bloc led by France endured contractions that lasted into 1935 and 1936. The single best predictor of the severity of the depression in different countries is how long they stayed on gold. The gold standard was a Midas touch that paralysed the world economy.

Real wages stayed high in countries on the gold standard. Macroeconomic policies to preserve the value of the currency reduced prices faster than wages, and real wages stayed high or even rose. Bank failures also were widespread in gold-standard countries, further depressing production. Both high real wages and bank failures show up as explanatory variables for low incomes around 1935, and the prevalence of financial crises in countries on gold suggests that a counterfactual with more rapid deflation and no devaluations would not have resulted in the maintenance of something close to full employment.

Complications

The influence of the gold standard determined the spread and the depth of the Great Depression, but the story has many dimensions not captured in this stark description. The literature can be contentious, although apparently competing views may represent elements in a more comprehensive view.

One view of the Great Depression sees it as an American contraction that was transmitted to the rest of the world. In *A Monetary History of the United States, 1867–1960* (1963), Milton Friedman and Anna Schwartz argued that the Federal Reserve System in the United States acted with such ineptness that it plunged the world into depression. They attributed this incompetence to the death of Benjamin Strong (president of the Federal Reserve Bank of New York) in 1928, and they describe several alternative monetary policies that they argue would have eased or even eliminated the economic contraction.

Even their story cannot separate the United States from the rest of the world, however. The Federal Reserve raised interest rates in October 1931 to defend the dollar, as noted above, even though the economy was contracting. Friedman and Schwartz characterized this action as an inept mistake, but they acknowledged the power of the gold standard to unite the financial community behind this perverse policy. This contractionary policy in the midst of rapid economic decline was the classic central bank reaction to a gold-standard crisis.

Charles Kindleberger put forward a more international explanation in *The World in Depression* (1986). He argued that the lack of central bank leadership in the operation of the restored gold standard was key to the spread of the Great Depression: the proposition summed up in the phrase ‘no longer London, not yet Washington’. The diminished financial status of Great Britain meant that London was unable to act as sole conductor of the international orchestra – or, in more modern terminology, to operate as the ‘hegemon’ – while the United States was not yet willing to take over this role despite the enormous improvement in its international economic standing.

Another factor which has been put forward as the primary explanation for the problems of the interwar period is the absence of international cooperation between the United States, Britain, France and Germany. Barry Eichengreen, in *Golden Fetters* (1992), identified this behaviour as a central feature of the period, manifest particularly in the attempt of each of the main powers to secure for itself a disproportionate share of the world’s limited stocks of monetary gold. Prior to the collapse of the gold standard in 1931 their non-cooperative behaviour involved the imposition of tight monetary policies not only by countries in deficit, but also by those – notably the United States and France – which were in surplus. This added to the deflationary pressures on the world economy and increased the vulnerability of the weak currencies, such as the pound and the mark, to speculative attack.

Recovery

The world began to recover from the contraction in 1933, when the United States and Germany

both abandoned the policies of the gold standard, but the Great Depression was far from over. Unemployment continued to be high in most countries, as indicated in Table 1. The world economy split up into competing currency and trading blocs, and domestic policies to combat the hardships of depression changed the role of government.

Unemployment continued to be high in most countries throughout the 1930s. Measures designed to help workers often perpetuated unemployment. The National Industrial Recovery Act of 1933 in the United States attempted to bring order to industries and income to workers by allowing industries to enforce codes of conduct that raised both prices and wages. Rising wages impeded the extension of employment, trading off the benefits to the unemployed for benefits to those working. Germany under the Nazis expanded government spending and, apparently, decreased unemployment dramatically. France and other members of the gold bloc continued to maintain contractionary policies in an effort to retain the convertibility of their currencies into gold. Only when France devalued in 1936 could its recovery begin.

Recovery, however slow and halting, did not approach the status quo ante. The world economy fragmented in the 1930s, and recovery took place within relatively isolated currency and trading blocs. The United States began the process of reducing world trade with the Smoot–Hawley tariff of 1930. The United Kingdom abandoned its tradition of free trade in 1932 in favour of protection for the British Commonwealth. Germany under the Nazis adopted a complex set of bilateral trading arrangements that reoriented its trade towards south-eastern Europe. International trade was much reduced, and international capital flows virtually disappeared.

Countries were changed internally as well. Governments became active in the economy as they attempted to reduce unemployment or ease the condition of the unemployed. Unions grew in many countries, helped both by legislation and by unemployment. Regulation grew as governments substituted direct controls for those of the market, and the world war that followed the Great Depression caused governments to take control even more firmly of their economies. The mixed economies and large governments that were typical of the last half of the 20th century were the legacy of the Great Depression and its aftermath.

It is not possible to separate the long-run effects of the Depression from those of the Nazis and the Second World War, but it is instructive to ask whether the Great Depression could have been avoided. There were indeed stresses on the world economy at the end of the 1920s, and the control mechanisms used in earlier times were not in good shape. The downturns in the United States and Germany would have produced a serious recession in the early 1930s in any case. The currency crises of 1931 then turned this recession into the Great Depression. If Germany and the United States had abandoned gold after Britain had chosen devaluation over further contraction, the world economy would have begun to recover two years earlier and before unbearable strain had been put on economic and political institutions.

Historians today debate how much freedom policymakers had in 1931. The German cabinet discussed devaluation after Britain left gold, but the memory of hyperinflation less than a decade before inhibited – if it did not preclude – an expansionary policy such as devaluation. The United States was not under the same economic pressure as Germany, but the Federal Reserve nonetheless raised interest rates sharply in late 1931 in response to gold outflows following Britain's devaluation. The Federal Reserve was following the dictates of the gold standard in actions that were applauded by the local financial community. It was a world tragedy – one that escalated from economics to politics and war – that the hold of the gold standard was so strong in the early 1930s that policymakers in the major economies chose to continue deflationary economic policies long after the need for expansionary measures was clear.

See Also

► Gold Standard
► Kindleberger, Charles P. (1910–2003)

Bibliography

Balderston, T. 1993. *The origins and course of the German economic crisis*. Berlin: Haude and Spener.

Bernanke, B. 1983. Nonmonetary effects of the financial crisis on the propagation of the great depression. *American Economic Review* 73: 257–276.

Bernanke, B. 1995. The macroeconomics of the great depression: A comparative approach. *Journal of Money, Credit, and Banking* 27: 1–28.

Bernanke, B., and H. James. 1991. The gold standard, deflation, and financial crisis in the great depression: An international comparison. In *Financial markets and financial crises*, ed. R.G. Hubbard. Chicago: University of Chicago Press.

Borchardt, K. 1991. *Perspectives on modern German economic history and policy*. Cambridge: Cambridge University Press.

Bordo, M., C. Goldin, and E.N. White. 1998. *The defining moment: The great depression and the American economy in the twentieth century*. Chicago: University of Chicago Press.

Calomiris, C.W. 1993. Financial factors in the great depression. *Journal of Economic Perspectives* 7(2): 61–85.

Calomiris, C.W., and J.R. Mason. 2003. Consequences of bank distress during the great depression. *American Economic Review* 93: 937–947.

Choudhri, E.U., and L.A. Kochin. 1980. The exchange rate and the international transmission of business cycle disturbances: Some evidence from the great depression. *Journal of Money, Credit, and Banking* 12: 565–574.

Eichengreen, B. 1992. *Golden fetters: The gold standard and the great depression*. New York: Oxford University Press.

Eichengreen, B. 2004. Still fettered after all these years. *Canadian Journal of Economics* 37: 1–27.

Eichengreen, B., and J. Sachs. 1985. Exchange rates and economic recovery in the 1930s. *Journal of Economic History* 12: 925–946.

Eichengreen, B., and P. Temin. 2000. The gold standard and the great depression. *Contemporary European History* 9: 183–207.

Feinstein, C.H., P. Temin, and G. Toniolo. 1997. *The European economy between the wars*. New York: Oxford University Press.

Field, A.J. 1984. Asset exchanges and the transactions demand for money, 1919–29. *American Economic Review* 74: 43–59.

Friedman, M., and A.J. Schwartz. 1963. *A monetary history of the United States, 1867–1960*. Princeton: Princeton University Press.

Hamilton, J.D. 1988. The role of the gold standard in propagating the great depression. *Contemporary Policy Issues* 6: 67–89.

James, H. 1986. *The German slump: Politics and economics, 1924–1936*. Oxford: Oxford University Press.

Kindleberger, C.P. 1986. *The world in depression, 1929–1939*, Rev. ed. Berkeley: University of California Press.

Margo, R.A. 1993. Employment and unemployment in the 1930s. *Journal of Economic Perspectives* 7(2): 41–59.

Mouré, K. 1991. *Managing the Franc Poincaré: Economic understanding and political constraint in French monetary policy, 1928–1936*. Cambridge: Cambridge University Press.

Romer, C.D. 1990. The great crash and the onset of the great depression. *Quarterly Journal of Economics* 105: 597–624.

Romer, C.D. 1993. The nation in depression. *Journal of Economic Perspectives* 7(2): 19–39.

Rothermund, D. 1996. *The global impact of the great depression, 1929–1939*. London: Routledge.

Temin, P. 1989. *Lessons from the great depression*. Cambridge, MA: MIT Press.

Temin, P. 1993. Transmission of the great depression. *Journal of Economic Perspectives* 7(2): 87–102.

Great Depression (Mechanisms)

Lee E. Ohanian

Abstract

This article summarizes the theoretical framework and the diagnostic procedures that economists use to construct and test theories of depressions and booms, and also summarizes recent applications of these procedures to well-known depressions.

Keywords

Balanced growth; Banking crises; Euler equations; Great Depression; Growth models; Optimal growth theory; Real business cycles

JEL Classifications

D4; D10

Depressions – prolonged periods in which output and employment fall 15 per cent or more below their long-run trend levels – are pathological. These episodes are particularly bizarre in economies like that of the United States, in which aggregate variables are almost always within a couple of percentage points of their long-run trend values (Leamer 2004). Economists have

long recognized that these abnormal episodes are strongly at variance with standard economic theory. For this reason, economists have not used equilibrium models, or, for that matter, any optimizing framework to investigate these episodes. And the reason why optimizing theories have been eschewed seems straightforward – what could equilibrium models tell us about episodes that appear to defy equilibrium reasoning? Prescott (2002) refers to the omission of theory from studies of depressions as a virtual 'taboo.'

The omission of theory from analysis of depressions and crises comes at a cost, as it limits the tools with which economists can investigate these pathologies and thus limits the extent to which we can understand them. Since the late 1990s, however, macroeconomists have begun to use theory to investigate depressions, with a focus on the application of optimal growth theory developed by Cass (1965) and Koopmans (1965). Obviously, the steady state growth path of the Cass–Koopmans model – by definition – fails to reproduce any depression episode. But economists are beginning to learn about the pathology of depressions as deviations from standard economic theory, much as a physician learns about illness by assessing deviations of a patient's vital signs from normality. The deviations of depressions from optimal growth theory are providing valuable diagnostic information that are used as the first step in developing and testing theories of depression. The remainder of this article describes these 'depression diagnostics', their application, and the new theories of depression that are being developed as a result of the use of these diagnostics.

The motivation behind the approach is that abnormal periods of macroeconomic activity – whether depressions or booms – lead to a proliferation of possible theories, and thus far there has been no systematic approach in the literature to shed light on which theories are the most promising. The diagnostic approach summarized here provides a simple method for identifying promising classes of theories. The idea of using optimal growth theory to diagnose depressions was initially used in Cole and Ohanian (1999), and the approach was further developed by Chari et al. (2002), Cole and Ohanian (2002), and in particular by Chari et al. (2007). The diagnostic method documents the deviations of the standard growth model, and then turns those deviations on their head to direct researchers to particularly promising classes of models.

To summarize the procedure, I begin with the following deterministic optimal growth model, which is given by:

$$\max \sum_{t=0}^{\infty} \beta^t \{\ln(C_t) + \varphi \ln(1 - L_t)\} N_t$$

where β is the household discount factor, C is consumption, L is hours worked, and N is population. The maximization is subject to the resource constraint, and the deterministic laws of motion for technology (X_t) which grows at the constant rate γ, and population (N_t) which grows at the constant rate n:

$$\begin{gathered} Y_t = F\left(K_t, (X_t L_t) + (1-\delta)K_t \geq C_t + K_{t+1}, K_0 \text{ given}\right. \\ X_t = (1+\gamma)^t X_0, X_0 \text{ given} \\ N_t = (1+n)^t N_0, N_0 \text{ given.} \end{gathered}$$

While this example uses logarithmic preferences, one can use other functional forms for utility. To induce stationarity, all variables are divided by population, and all variables that grow along the steady state growth path are detrended by the exogenous productivity factor $(1 + \gamma)^t$. The stationary, per-capita variables are denoted with lower-case variables. Standard dynamic programming techniques can be used to solve for the first-order conditions. The equations that characterize the planner's optimum, and that form the basis of the diagnostic procedure, are given by:

$$\begin{gathered} \varphi/(1-l_t) = F_{lt}/c_t \\ (1+\gamma)(1+n)c_{t+1}/c_t = \beta[F_{kt+1} + 1 - \delta] \\ y_t = c_t + (1+\gamma)(1+n)k_{t+1} - (1-\delta)k_t \\ y_t = F(k_t, x_0 l_t) \end{gathered}$$

The first of the four equations listed above governs the household's allocation of time between market and non-market activities. The second equation, often called the Euler equation,

governs the household's allocation of income between consumption and savings. The third equation is the resource constraint, and the fourth equation is the production function. Given parameter values, a functional form for the technology, and time series observations on output, consumption, labour, and investment, we construct the following percentage deviations between theory and data that will form the basis of the diagnostics:

$$\varepsilon_{lt} = \{\varphi/(1-l_t)/(F_{lt}/c_t)\} - 1$$
$$\varepsilon_{kt} = \{(c_{t+1}/c_t)/(\beta[F_{kt+1} + 1 - \delta])\} - 1$$
$$\varepsilon_{yt} = \frac{y_t - c_t - i_t}{y^* - c^* - i^*} - 1,$$

where starred variables are steady state values, and

$$\varepsilon_{zt} = [y_t/F(kt, x_0 l_t)] - 1$$

I denote these as the labour deviation, (ε_{lt}) the Euler deviation (ε_{kt}), the National Income and Product Accounts (NIPA) deviation (ε_{yt}), and the productivity deviation (ε_{zt}). With the exception of the NIPA equation, each of these deviations is constructed as the percentage difference between the left-hand side and the right-hand side of each equation. The NIPA deviation is normalized relative to a steady state value. Note that, along the steady state growth path, all of these deviations are equal to zero by construction.

The next step is to choose parameter values and functional forms. Regarding parameterization, it is often convenient to choose values so that the deviations are equal to zero immediately before the episode of interest. Thus, given values for consumption, hours worked, and the technology, the parameter value for φ is chosen so that the $\varepsilon_{lt} = 0$ prior to the episode. Similarly, given values for consumption, capital, hours worked, and the depreciation rate, the parameter value for β is chosen so that (ε_{kt}) is zero prior to the period of interest. Regarding functional forms, the application here uses log preferences over consumption and leisure, though other forms can certainly be used. For the production function, it is common to use Cobb-Douglas.

For heuristic purposes, I have described the procedure in a deterministic economy. The extension to a stochastic economy is fairly straight forward, and is presented in detail in Chari et al. (2007). In summary, the procedure is modified for stochastic environments so that the measured deviations are modelled as a vector stochastic process, which is typically specified as a vector autoregression (VAR) and which for simplicity is represented here as a first-order process. The vector W is a 4×1 vector containing the four deviations previously defined above. The VAR is given by:

$$W_t = \varphi W_{t-1} + \varepsilon_t, E(\varepsilon\varepsilon') = \Omega$$

The labour, NIPA, and productivity deviation are measured exactly the same way as in the deterministic case. Measuring the Euler deviation, however, requires evaluating the expectation of the right-hand side of the Euler equation. This can be accomplished by log-linearizing the planner's conditions, and then solving the resulting linear system. The linearized model can then be used to forecast the right-hand side of the Euler equation, which implicitly defines the stochastic Euler deviation.

Chari et al. (2007) show that, given this procedure of measuring the deviations, the model economy is able to reproduce (up to numerical solution error) any observed sequences of fluctuations of the endogenous variables, given the sequences of these deviations and an initial value of the capital stock, population, and technology. Consequently, *any* fluctuation of an actual economy from its balanced growth path value is *entirely* accounted for by one or more of these deviations within the growth model. This insight transforms the growth model into an accounting framework, and as such the procedure can document the relative importance of each deviation for understanding fluctuations. To do this, the investigator calculates the solution of the model using just one deviation at a time, or a subset of

the four deviations. Chari et al. (2007) conducted this type of analysis to show that the labour deviation and the productivity deviation account for virtually all of the Great Depression through 1933.

It is important to recognize that the interpretation of the deviations at this stage of analysis differs from that in the business cycle literature. For example, real business cycle models typically identify the Solow residual, or some variant of that residual, as a primitive shock. The diagnostic framework summarized here does not necessarily place this type of identifying interpretation on these deviations. Rather, the focus here is to provide clues to researchers for the class of models to consider. In the case of a productivity deviation, the key point of the diagnostic is that it informs researchers that a successful theory will be one that can be mapped into a growth model with this feature. A shift in the aggregate technology set is one interpretation of this deviation, but there are other interpretations as well, including theories based on the mismatch of resources across plants which impacts the Solow residual (see Restuccia and Rogerson 2007), a distortion in relative prices that leads firms to shift their input mix, which also impacts the Solow residual(see Chari et al. 2007), or changes in government regulations (see Hansen and Prescott 1993).

To concretely illustrate the use of this diagnostic approach further, suppose an investigator calculated these deviations, and found they were all roughly zero with the exception of the labour deviation. This information narrows the class of theories so that it would be admissible to include only those that feature some mechanism that changes the rate at which households value their time with respect to the measured wage, but leaves all other margins within the growth model unchanged. Models in this class include those with time-varying taxes on labour income, time-varying subsidies to non-market time, such as changes in unemployment benefits, changes in the incentive to accumulate human capital, and changes in union or monopoly power. The diagnostics presented in this hypothetical example also exclude several classes of models, including models of productivity shocks, models with government spending shocks, and models with time-varying taxes on capital income or investment, as the margins on which these factors operate are not distorted.

This example of a large labour distortion, but no other large distortions, not only illustrates how the method can be applied, but also happens to be the outcome of the procedure when the tools are applied to two well known and puzzling depressions, in the United States between 1933 and 1939 and the United Kingdom between 1921 and 1929. These episodes have long been considered puzzling because of their long duration, and because standard economic fundamentals were reasonably healthy during this period. In particular, the US money supply grew quickly after 1933, productivity growth was rapid after 1933, and there were no bank runs. All of these factors should have fostered a rapid recovery in the United States, yet hours worked in the United States recovered very little after the 1933 trough. In the United Kingdom, productivity growth was at its trend level during the 1920s, and the economy should have been poised for a significant post-First World War recovery.

During both of these episodes ε_{lt} was very large, but other deviations were small. Specifically, the marginal rate of substitution fell well below the wage in both of these depressions. In the United States, the real wage was as much as 100% above the marginal rate of substitution, while total factor productivity was near trend in both episodes, and the intertemporal consumption-savings Euler equation was undistorted. The diagnostic thus establishes that the key depressing factors in these episodes severely affected the labour market but did not depress productivity, nor did they distort the household's intertemporal first-order condition governing the consumption savings decision.

Cole and Ohanian (1999, p. 6) used this diagnostic information to focus on theories which could substantially distort the household's time

allocation decision. Perhaps the most obvious factor that distorts this decision is changes in labour and consumption tax rates, as these taxes change the incentive to work and thus impact this first-order condition. The authors ruled this factor out based on empirical grounds, because changes in labour and consumption taxes were relatively small during this period in both countries.

For the United States, the next factor they considered was government policies that impact the labour market, and major labour market policies were adopted in both countries just prior to these episodes. In the United States, President Roosevelt introduced the National Industrial Recovery Act in 1933 that permitted industrial firms to cartelize and raise prices provided that they also raised wages. Relative prices and real wages in these sectors jumped significantly after the adoption of these policies, and employment and output remained low throughout the 1930s. This led Cole and Ohanian (2004) to develop a dynamic insider–outsider model in which firms were able to collude provided they reached a wage agreement with their workers. The model accounted for about 60% of the post-1933 depression, and was consistent with the behaviour of wages and prices in both the industrial and non-industrial sectors of the economy. The insider–outsider model developed by Cole and Ohanian maps into a standard growth model with a large labour deviation, but the other deviations are roughly zero.

For the United Kingdom, Cole and Ohanian (2002), following work by Benjamin and Kochin (1974), noted that the United Kingdom adopted a very generous unemployment policy after First World War. Initially, benefits were available to those who worked for only one day in total, could be received indefinitely, and provided generous payments. At one point, the benefits paid were equal to about four per cent of GNP. Cole and Ohanian developed a model in which the policy reduced hours worked by about 15 per cent and distorted the household's first-order condition governing time allocation, but did not affect the other margins in the growth model.

In both of these episodes, the substantive finding from the diagnostic procedure led to the development of theories in which government policies distorted labour markets and significantly reduced hours worked, but did not significantly distort the incentive to save. The most provocative issue that has arisen in this diagnostic literature regards the importance of distortions to the capital market as a source of depression in the United States in the 1930s. Bernanke (1983), in a very influential paper, argued that banking panics led to a longer and deeper depression, using regression analysis that demonstrated that banking variables were statistically significant in an output equation that also included a measure of monetary shocks. Chari et al. (2007) show that some optimizing models that feature the financial intermediation channel emphasized by Bernanke, including Carlstrom and Fuerst (1997), map into a growth model with a substantial Euler equation deviation. However, the empirical Euler deviation is small during this period, leading Chari, Kehoe, and McGrattan to conclude that financial frictions theories that operate through the specific channel emphasized in Carlstrom and Fuerst are not quantitatively important for the Great Depression. This finding was a surprise to many economists, and is leading to new research in this area (see Christiano and Davis 2007; Primaceri et al. 2006).

It is likely that this procedure will be used not only in business cycles, but in a variety of applications. Lu (2007) uses the procedure to study Taiwan since the 1950s, and finds a very large Euler deviation. Given this finding, and the fact that the spread between deposit and lending rates in Taiwan declines from about 12 per cent in 1950 to about two per cent in 2003, she develops a model of technological advances in financial intermediation efficiency that can account for the decline in the saving–lending spread and the Euler deviation, and finds that this development led to a significant increase in Taiwanese per-capita income.

See Also

- ▶ Great Depression
- ▶ Great Depression, Monetary and Financial Forces in
- ▶ Growth and Cycles

Bibliography

Benjamin, D., and L. Kochin. 1974. Searching for an explanation of unemployment in interwar Britain. *Journal of Political Economy* 87: 441–478.

Bernanke, B. 1983. Nonmonetary effects of the financial crisis in the propagation of the great depression. *American Economic Review* 73: 257–276.

Carlstrom, C., and T. Fuerst. 1997. Agency costs, net worth, and business fluctuations: A computable general equilibrium analysis. *American Economic Review* 87: 893–910.

Cass, D. 1965. Optimum growth in an aggregative model of capital accumulation. *Review of Economic Studies* 32: 233–240.

Chari, V.V., P. Kehoe, and E. McGrattan. 2002. Accounting for the great depression. *American Economic Review* 92: 22–27.

Chari, V.V., P. Kehoe, and E. McGrattan. 2007. Business cycle accounting. *Econometrica* 75: 781–836.

Christiano, L., and Davis, J. 2007. *Two flaws in business cycle accounting*. Discussion paper, Department of Economics, Northwestern University.

Cole, H., and L.E. Ohanian. 1999. The great depression in the United States from a neoclassical perspective. *Federal Reserve Bank of Minneapolis Quarterly Review* 23(1): 2–24.

Cole, H., and L.E. Ohanian. 2002. The U.S. and U.K. great depressions through the lens of neoclassical growth theory. *American Economic Review* 92: 28–32.

Cole, H., and L.E. Ohanian. 2004. New Deal policies and the persistence of the Great Depression. *Journal of Political Economy* 112: 779–816.

Hansen, G., and E. Prescott. 1993. Did technology shocks cause the 1990–91 recession? *American Economic Review* 83: 280–286.

Koopmans, T. 1965. *On the concept of optimal economic growth*. New York: Rand McNally.

Leamer, E. 2004. The truth about GDP growth. *Harvard Business Review,* October, 24.

Lu, S.-S. 2007. *Understanding growth miracles: The case of Taiwan*. Ph.D. thesis, UCLA.

Prescott, E. 2002. Prosperity and depression. *American Economic Review* 92: 1–15.

Primaceri, G., E. Schaumburg, and A. Tambalotti. 2006. *Intertemporal disturbances*, Working Paper No. 12243. Cambridge, MA: NBER.

Restuccia, D., and R. Rogerson. 2007. *Policy distortions and aggregate productivity with heterogeneous plants*, Working paper. Cambridge, MA: Department of Economics, University of Toronto.

Great Depression, Monetary and Financial Forces In

Satyajit Chatterjee and P. Dean Corbae

Abstract

We survey papers that seek model-based answers to the following questions regarding the Great Depression. What caused the worldwide collapse in output from 1929 to 1933? Why was the recovery from the trough of 1933 so protracted for the United States? How costly are Depression-like episodes in terms of welfare? Was the decline in output preventable? The papers point to: an important, but not exclusive, role of monetary factors in causing the decline; counterproductive labour market interventions in making the recovery slow; uninsured risk of unemployment in making Depression-like episodes costly; timely provision of liquidity as a preventive policy.

Keywords

Confidence; Debt-deflation hypothesis; Depressions; Dynamic stochastic general equilibrium (DSGE) model; Financial intermediation; Gold standard; Great Depression; Liquidity preference; Monetary and financial forces in the Great Depression; Monetary base; Money multiplier; Multiple equilibria; Sticky wages; Total factor productivity

JEL Classifications

D4; D10

What caused the worldwide collapse in output from 1929 to 1933? Why was the recovery from

the trough of 1933 so protracted for the United States? How costly was the decline in terms of welfare? Was the decline preventable? These are some of the questions that have motivated economists to study the Great Depression.

Cole and Ohanian (1999) document that US per capita GNP fell 38 per cent below its long-run trend path (of two per cent per annum growth) from 1929 to 1933. Real per capita non-durables consumption fell nearly 30 per cent, durables consumption fell over 55 per cent, and business investment fell nearly 80 per cent. On the input side, total employment fell 24 per cent and total factor productivity (TFP) fell 14 per cent. On the nominal and financial side, the GNP deflator fell 24 per cent; per capita M1 (currency plus deposits) fell 30 per cent; M1 velocity fell 32 per cent; the per capita monetary base rose 9 per cent; the currency–deposit ratio rose over 160 per cent (Friedman and Schwartz 1963, Table B3); the loan–deposit ratio fell 30 per cent (Bernanke 1983, Table 1); and *ex post* real commercial paper rates rose from six per cent in 1929 to a peak of 13.8 per cent in 1932.

What caused the Depression? For the United States, Friedman and Schwartz (1963, p. 300) argued that it was the decline in the stock of M1 – a consequence of Fed tightening and of a fall in the money multiplier induced by banking panics. According to Eichengreen (1992), international adherence to the gold standard transmitted the US monetary contraction to other industrialized countries.

Specifically, high interest rates and low prices in the United States attracted foreign inflows of gold (in 1932 the United States and France held over 70 per cent of the world gold reserves), which the Fed largely sterilized (that is, sold domestic government debt and bought money). The outflow of gold from foreign countries implied that gold-backed money supplies of those countries had to decline in order to meet their cover ratios. Further evidence (see Bernanke and James 1991, Table 4) of the importance of the gold standard in transmitting the contraction comes from the experience of countries like Britain, which suspended the gold standard in 1931 and recovered by 1932; from Spain, which never was on it and had a much less severe contraction than those on the gold standard; and from France, which was one of the last major countries to leave it and still faced declining industrial production past the 1933 trough. As Bernanke (1995, p. 3) puts it: 'The new gold-standard research allows us to assert with considerable confidence that *monetary factors played an important causal role*, both in the worldwide decline in prices and output and in their eventual recovery.'

However, much of this evidence is problematic in that it is in the nature of correlations between *endogenous* variables – a fact that makes it challenging to establish causality. Did the decline in M1 *cause* the decline in aggregate output or – as Temin (1976) argued early on – did M1 and aggregate output decline in response to some other common shock? If the 'monetary-cum-exchange- rate-policy' explanation is indeed correct, we ought to be able to demonstrate its correctness in a reasonably calibrated, dynamic stochastic general equilibrium (DSGE) model. To paraphrase Lucas (1993, p. 271): 'If we know what a depression is, we ought to be able to *make* one.' The challenge of 'making' a depression has been taken up by various researchers and constitutes a noteworthy recent development in depression research.

The conventional explanation of why money affected output is sticky nominal wages – goods prices fell as a result of the monetary contraction but nominal wages adjusted slowly and the ensuing increase in the real wage depressed the demand for labour. One significant contribution to evaluating this conventional explanation is by Bordo et al. (2000). They calibrate a one-sector stochastic macro model with four-quarter nominal wage rigidity and find that 70 per cent of the output decline from 1929 to 1933 can be accounted for by feeding in the negative innovations to the actual M1 money supply process during that period.

Although the findings of Bordo, Erceg and Evans are striking, there are some unresolved issues. One is that the real-wage rise in the model was chosen to mimic the actual real-wage rise in the manufacturing sector while there is some indirect evidence that non-manufacturing

real wages actually fell during the 1929–33 downturn. Cole and Ohanian (2000) re-examine the sticky-wage hypothesis in a multisector model and find much less support for it.

A second unresolved issue is that Bordo, Erceg and Evans do not take into account the evidence on aggregate labour productivity and TFP, both of which declined between 1929 and 1933. Ohanian (2002) argues that only about a third of the decline in labour productivity and/or TFP can be plausibly accounted for by mismeasurement of factor inputs. By itself, a decline in TFP could account for a substantial fall in aggregate output, consumption and investment. Unless a decline in TFP can be viewed as an endogenous response to the monetary shock (through, for example, aggregate increasing returns), the decline leaves less scope for a purely monetary explanation. Using a DSGE model where money is non-neutral due to imperfect information, Cole et al. (2005) show that the decline in M1 accounts for only one-third of the decline in output from 1929 to 1933, while the effect of an exogenous decline in TFP accounts for two-thirds. They use a misperceptions model of monetary non-neutrality because such a model generates less of a counterfactual movement in labour productivity than a model with nominal wage rigidities.

Sticky wages and monetary misperceptions are not the only mechanisms through which money can affect real output. Irving Fisher (1933) pointed out that the unanticipated fall in prices during 1929–33 led to bankruptcies because it increased the real value of nominal debt of households, firms, and financial intermediaries. This 'debt-deflation' hypothesis was analysed by Mishkin (1978) for households and formalized by Bernanke and Gertler (1989) for firms. More generally, Bernanke (1983) argued that the reduction in borrower net worth increased the cost of obtaining external finance, while bank failures and tightened credit standards hampered the efficient allocation of capital. However, a quantitative DSGE model featuring this mechanism has yet to be implemented for the Great Depression. Such a model holds out the promise of explaining some portion of the puzzling decline in TFP during 1929–33 as an endogenous response to a misallocation of capital.

One of the most striking facts of the Depression was the reduction in the money multiplier from 1929 to 1933 associated with the flight from bank deposits to currency. Cooper and Corbae (2002) construct a model in which households have the option of saving in the form of currency or bank deposits, and in which bank deposits ultimately fund working capital for businesses. Because of increasing returns in the intermediation technology associated with fixed verification costs, their model admits multiple equilibria. In the good equilibrium the return on bank deposits is high, households hold small amounts of currency, and output is high. In the bad equilibrium, the return on bank deposits is low, households substitute into currency, and output is low. A shift from the good to the bad equilibrium replicates many of the salient nominal changes that occurred between 1929 and 1933. Although not quantitative, their work formalizes the idea that output, credit and money supply responded negatively to a loss in confidence – much as Irving Fisher (1933, p. 343) suggested it did.

Why was the recovery from the trough of 1933 so protracted for the United States? As noted by Cole and Ohanian (1999), aggregate US output was still below trend in 1939. The answer cannot be the gold standard or M1 because the United States left the gold standard in 1933 and the US money stock recovered rapidly thereafter. One explanation offered is that the National Industrial Recovery Act (NIRA) encouraged businesses to accept high real wages of industrial workers. Cole and Ohanian (2004) embed labour bargaining into a DSGE model and quantitatively explore the effect of the NIRA, giving more weight to workers in the bargaining process post 1933. Their model is reasonably successful in producing a slow recovery. Adverse labour market interventions also appear to have played a role in other industrialized countries such as Germany, France, the UK and Italy (Kehoe and Prescott 2002).

How costly was the Depression in terms of welfare? Real per capita consumption of non-durables fell 30 per cent in the United States

but it is not known how this decline was distributed across households. Chatterjee and Corbae (2007) analyse how households that can self-insure against uninsured earnings losses would fare through a depression. They found that the welfare cost of living in a world with a small likelihood of a Depression-like event is quite large – somewhere between one and seven per cent of consumption in perpetuity depending on the completeness of asset markets. Much of this cost is associated with the increased variability of individual consumption streams.

Was the Depression preventable? First, if the 'monetary-cum-exchange-rate- policy' explanation is correct, the right monetary policy could have prevented the decline. Christiano et al. (2003) estimate a DSGE model with many shocks but find that a liquidity preference shock inducing households to hold currency instead of deposits played the most important role in the contraction phase of the Depression. They then specify a policy rule that raises the monetary base as a function of liquidity shocks, and run a counterfactual experiment where they find that output would have declined only six per cent if such a reaction function had been in place. Second, if a portion of the decline in output was the result of a banking collapse stemming from a shock to confidence, then – as shown by Cooper and Corbae (2002) – an announcement by the monetary authority that it stands ready to supply liquidity to the banking system might have moderated the decline. Finally, with regard to the slow recovery in the United States, the only credible explanation offered is adverse labour market intervention. If this explanation is correct, we know what *not* to do to prolong a severe decline in output.

See Also

- ▶ Great Depression
- ▶ Great Depression (Mechanisms)
- ▶ Monetary Business Cycle Models (Sticky Prices and Wages)
- ▶ Real Business Cycles

Bibliography

Bernanke, B. 1983. Nonmonetary effects of the financial crisis in the propagation of the Great Depression. *American Economic Review* 73: 257–276.

Bernanke, B. 1995. The macroeconomics of the Great Depression: A comparative approach. *Journal of Money, Credit and Banking* 27: 1–28.

Bernanke, B., and M. Gertler. 1989. Agency costs, net worth, and economic performance. *American Economic Review* 79: 14–31.

Bernanke, B., and H. James. 1991. The gold standard, deflation, and financial crisis in the Great Depression: An international comparison. In *Financial markets and financial crises*, ed. R. Glenn Hubbard. Chicago: University of Chicago Press.

Bordo, M., C. Erceg, and C. Evans. 2000. Money, sticky wages, and the Great Depression. *American Economic Review* 90: 1447–1463.

Chatterjee, S., and D. Corbae. 2007. On the aggregate welfare cost of Great Depression unemployment. *Journal of Monetary Economics* 54: 1529–1544.

Christiano, L., R. Motto, and M. Rostagno. 2003. The Great Depression and the Friedman–Schwartz hypothesis. *Journal of Money, Credit, and Banking* 35: 1119–1197.

Cole, H., and L. Ohanian. 1999. The Great Depression in the United States from a neoclassical perspective. *Federal Reserve Bank of Minneapolis Quarterly Review* 23(1): 25–31.

Cole, H., and L. Ohanian. 2000. Re-examining the contributions of money and banking shocks to the U.S. Great Depression. *NBER Macroeconomics Annual* 15(1): 183–227.

Cole, H., and L. Ohanian. 2004. New Deal policies and the persistence of the Great Depression: A general equilibrium analysis. *Journal of Political Economy* 112: 779–816.

Cole, H. Ohanian, L., and Leung, R. 2005. Deflation and the international Great Depression: A productivity puzzle, Working Paper No. 11237. Cambridge, MA: NBER.

Cooper, R., and D. Corbae. 2002. Financial collapse: A lesson from the Great Depression. *Journal of Economic Theory* 107: 159–190.

Eichengreen, B. 1992. *Golden Fetters: The gold standard and the Great Depression 1919–1939*. New York: Oxford University Press.

Fisher, I. 1933. The debt–deflation theory of Great Depressions. *Econometrica* 1: 337–357.

Friedman, M., and A. Schwartz. 1963. *A monetary history of the United States, 1867–1960*. Princeton: Princeton University Press.

Kehoe, T., and E.C. Prescott. 2002. Great Depressions of the twentieth century. *Special Issue of the Review of Economic Dynamics* 5(1): 1–18.

Lucas, R. 1993. Making a miracle. *Econometrica* 61: 251–272.

Mishkin, F. 1978. The household balance sheet and the Great Depression. *Journal of Economic History* 38: 918–937.
Ohanian, L. 2002. Why did productivity fall so much during the Great Depression? *Federal Reserve Bank of Minneapolis Quarterly Review* 26(2): 12–17.
Temin, P. 1976. *Did monetary forces cause the Great Depression?* New York: W.W. Norton and Company.

Great Divide

Erik Berglöf

Keywords
Commitment; Compensation; Crowding out; Fiscal responsibility; Great Divide; Policy reform; Privatization; Rule of law; Transition

JEL Classifications
P3

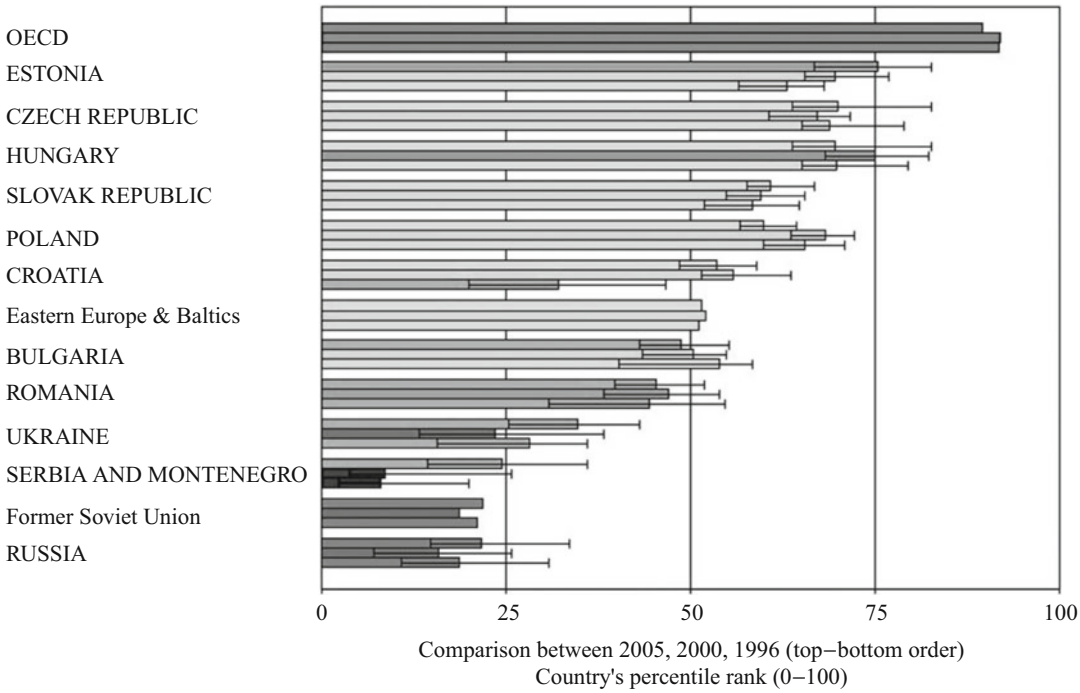
By the end of the first decade of transition in 2000, the Great Divide was visible in almost every measure of economic performance: gross domestic product (GDP) growth, investment, government finances, growth in inequality, general institutional infrastructure, and in measures of financial development. Gradually growth has picked up and macro-stability improved in the Commonwealth of Independent States countries, in large part thanks to favourable terms-of-trade changes as a result of price increases in energy and other raw materials. As indicated by Fig. 1, institutional development has also progressed in most of these countries but still lags significantly on most dimensions, and there are examples of institutional regression, particularly in political institutions.

The expression 'Great Divide' refers to the differences in institutional development apparent in the first decade of transition among the countries of central and eastern Europe and former Soviet Union following the collapse of Communism (Berglöf and Bolton 2002). Figure 1 compares these countries in 1996, 2000 and 2005 using one measure of rule of law derived through questionnaires to businesses operating in a broad range of countries.

The Great Divide does not primarily refer to the depth and duration of the initial transitional recession, but rather to the more long-term institutional backwardness observed in parts of the region. Some countries (for example, Estonia, Latvia and Lithuania) had a deeper recession, but a less protracted turnaround accompanied by rapid institutional transformation. Other countries (for example, Belarus, Turkmenistan, and Uzbekistan) did not attempt full-scale reform, and thus experienced neither initial decline nor substantial development in their institutions. The non-European transition countries China and Vietnam exhibit yet another pattern where broad, if incremental and partial, economic reform did not result in initial output decline.

The Great Divide represents one of the key puzzles of transition in that it cannot be immediately traced to the policies pursued. In fact, differences in policies among the more successful countries in Central and Eastern Europe were often more pronounced than those between some of these countries and those of former Soviet Union. A prominent, but by no means the only, example would be privatization policies, where the Czech model had more in common with that of Russia than with that of Poland.

Another, perhaps less puzzling, observation is the remarkable institutional convergence of economic systems in Central and Eastern Europe despite the diversity in terms of policies pursued. The emerging model of ownership and control of large firms is one with an owner with a large controlling share and a strong presence in day-to-day management of the firm. The financial systems are strongly dominated by commercial banks, increasingly foreign-owned, whereas stock markets on the whole remain volatile and illiquid. This convergence has been taking place even though the countries differ markedly in terms of



Great Divide, Fig. 1 Rule of law (world, 2005) (Source: Kaufmann et al. 2006)

policies when it comes to areas such as enterprise and bank privatization, bank recapitalization, stock market policies, and entry and exit of firms.

Both these observations, the emergence and persistence of the Great Divide and the convergence of economic systems in the front-runner countries in central and eastern European countries, suggest that initial conditions matter greatly for institutional development and economic growth. The relative importance of different initial conditions has been estimated by a large number of studies, but the influences of individual factors are often hard to disentangle. Broadly speaking, the Soviet legacy (the degree of integration into the economic and political system of the Soviet Union) and the prospect of membership in the European Union stand out as key in shaping the development of economic and political institutions.

Another key to understanding the origin of the Great Divide is to look at when and under what conditions it emerged. Typically, the differences in institutional development first became visible when the governments were faced with demands from different groups to be compensated for the adjustments in relative prices following pricing reforms. The ability of governments of transition economies to achieve fiscal and monetary responsibility, together with a commitment to refrain from bailing out failing banks or loss-making enterprises, determined whether economic and financial development took off.

Fiscal responsibility promotes both financial development and economic growth through two important channels: it limits the extent of crowding out of private investment by government borrowing and it makes credible the commitment of the government to maintain the macro-stability essential for private investment. In addition, it provides some guarantees that the returns from investment are not going to be taxed away in the future by excessively profligate governments desperately seeking tax revenues where they can find them. Of course, specific initial conditions and underlying country characteristics facilitate the emergence of fiscally sound governments capable of enforcing the rule of law.

See Also

- ▶ Emerging Markets
- ▶ Financial Structure and Economic Development
- ▶ Privatization
- ▶ Transition and Institutions

Bibliography

Berglöf, E., and P. Bolton. 2002. The great divide and beyond – Financial architecture in transition. *Journal of Economic Perspectives* 16(1): 77–100.

Kaufmann, D., A. Kraay, and M. Mastruzzi. 2006. *Governance matters V: Governance indicators for 1996–2005*. Washington, DC: World Bank.

Greek Crisis in Perspective: Causes, Illusions and Failures

Nicos Christodoulakis

Abstract

Soon after the 2008 global crisis, the euro faced its toughest challenge since its introduction as several of the participating Member States faced unprecedented financial problems. Greece was the most severe case, requiring intervention from the EU and IMF to stabilise its economy and repay debt obligations. This article explains the debt process in Greece from the 1980s to date, and describes its main causes and episodes. It also assesses the impact of the IMF/EU austerity programmes and shows that its failure to control recession inhibited the prospects of debt stabilisation. Five years after the bailout agreement, Greece is in a deep contraction, with socially explosive unemployment rates, while public debt is alarmingly higher than the level that triggered the crisis. An alternative scenario is discussed, showing that stabilisation can become more effective and realistic if recession is tackled first and reforms follow on a steadier path.

Keywords

Debt; Fiscal deficits; External balances; Crisis; Eurozone; Greece

JEL Classifications

H60; H61

Introduction

The elections that took place in Greece in January 2015 marked a wholly new phase of the socio-economic crisis that has been ravaging the country since 2010. The party of the Radical Left Coalition (Syriza) won the elections and subsequently formed a government in alliance with the nationalist eurosceptic party of Independent Greeks. The main priorities were announced to be the dismantling of austerity policies, the reinstatement of civil servants who were laid off in previous years as part of the fiscal contraction and a radical renegotiation of the terms of debt repayments to lenders. This led to protracted negotiations with the International Monetary Fund (IMF), the European Union (EU) and the European Central Bank (ECB), which had co-financed public deficits after Greece sought a bailout agreement in May 2010.

As the conditions of the bailout programme are now in jeopardy, the proper continuation of loans is put in question while Greece is withheld from participating in the programme of quantitative easing (QE) that is being implemented by the ECB. A great deal of uncertainty has since prevailed in Greece, causing further capital flight, a complete abstention from new investment and increasing liquidity pressure. No wonder that the risk of Greece exiting the eurozone is rising (the so-called 'Grexit'), adding further nervousness and anxiety across the spectrum of economic activity.

It all began in the aftermath of the global financial crisis of 2008, when a number of eurozone countries were engulfed in a spiral of rising public deficits and explosive borrowing costs that eventually drove them out of international markets. Greece was by far the most perilous case, with a double-digit fiscal deficit, an accelerating public debt which

in GDP terms was twice as much the eurozone average and an external deficit near US$5,000 per capita in 2008, one of the largest worldwide. No wonder that Greece was the first to seek the bailout assistance, followed later by Portugal in the end of 2010, Ireland in 2011 and Cyprus in 2013. All three countries have either terminated the conditionality programme or are in the last stage of doing so. Greece, however, is still engulfed in the programme, with no visible termination date.

In fact, the present threat of Grexit is not new, but has occurred a number of times over the past five years, whenever the programme was facing serious difficulties in its implementation. The first time was in the summer of 2011, when domestic uncertainties multiplied at such a rate that the possibility of Greece exiting the eurozone was widely discussed, either as a punishment mechanism from abroad for Greeks not accepting the pains of adjustment, or as a quick fix from within to avoid such pains for good.

Alarmed by those developments, two subsequent EU summits, held respectively in July and October 2011, decided that the Memorandum agreement should be substantially broadened to include a radical debt reduction, a second round of bailout loans by IMF and the EU and a generous release of European structural funds to assist the real economy. The agreement was conditional on being approved by the national Parliaments of the lender states as well as by the European Parliament. Finally, the conditionality of the Memorandum was approved by the Greek parliament in February 2012 and the debt-cutting process was concluded in May that year. However, most of the envisaged measures were delayed for the third quarter of the year, as two rounds of elections took place to provide new legitimacy for carrying on the programme and implementing reforms.

The prolonged electoral uncertainty meant that most of the adjustment measures were weakened or postponed, leading to new tensions over Greece's determination to implement the programme. As the first round of elections was inconclusive, the threat of Grexit spread all over again, before a new coalition government was finally formed in June by parties vowing to apply all policies deemed necessary for the country to remain in the eurozone. During its two and a half years in power, the coalition government achieved some visible progress in harnessing both the public and the external deficit, though public debt continued to grow out of control. On the other hand, the real economy continued to face an unprecedented recession, unemployment was rocketing and social despair undermined the implementation of market reforms. Exports remained anaemic, despite extensive cuts in wages and salaries.

After their massive defeat in the European elections in May 2014, the coalition parties sought for some relaxation of the stringent austerity time frame and a gradual emancipation from the policy supervision by the IMF and the European authorities. But this attempt failed, and the huge swing of votes to the Radical Left party in the new elections led to yet another impasse. Five years after the bailout Memorandum was signed, the situation remains ambivalent and far from been easy to stabilize. On the other hand, the European institutions have – in comparison with the improvisations in handling the 2008 crisis – become a lot more efficient to handle the credit shortages and mitigate the consequences of recession. The quantitative easing of the ECB, the new investment plan by the EU, the fiscal compact to monitor economic policy in member states and the pursuit of more coordination in the functioning of European banks are the four new pillars of stability and vigilance in the monetary union. Had such mechanisms been in existence right after the global crisis, several of its ramifications would have been milder and some of the stabilisation failures could have been avoided. From this point of view, it would be an irony of history if Greece now abolishes the capacity to adjust, and thus misses the opportunities for more growth and stability.

This combination of domestic fatigue and European re-invigoration makes the Greek problem an unusually interesting case for analysis, not only for understanding its origins and causes, but also for devising a realistic strategy to resolve it.

The purpose of the present article is twofold: first to provide a historical account of debt

accumulation, identify the main difficulties of fiscal stabilisation and explain the factors that led to the present crisis or failed to thwart it; and second, to assess the main reasons for at least thus far missing the targets set by the Memorandum agreement and the need for encompassing a growth strategy in order to make reforms acceptable and more effective to achieve debt sustainability in the longer run.

The next section describes the main episodes of debt escalation in the 1980s. Following that we consider the stabilisation effort on the way to EMU and the toxic combination of fiscal irresponsibility, external deficits and political indecision during the more recent period that led to the bail-out. Some recurrent facts on fiscal policies that repeatedly hinder stabilisation and growth are then described, followed by an attempt for an *ex post* assessment of the policies conditioned by the Memorandum agreement to correct the economy. The penultimate section argues why exiting the eurozone should not be an option for Greece. An alternative scenario based on more realistic (i.e. smaller) fiscal surpluses in exchange for higher growth is shown to be more credible in achieving fiscal consolidation and stabilising the debt over the medium term. We then conclude with the need to fight the current recession as the only way for Greece to regain social coherence and debt sustainability in the new landscape of the eurozone.

In its present form, the article is an update of the original published three years ago in the same series. But apart from adding more recent data on various economic indicators and briefly describing the socio-economic developments in between, none of the original analysis and policy assessments has been retroactively adapted. The reader is thus able to judge the validity of the early assertions regarding the viability of the programme and the need to urgently revise it. In fact, the current version reveals that the only deviations from the original predictions are for the worse: for example, growth is further below, and public debt is further higher, than the pessimistic trajectories assumed three years ago. Obviously, it is hard to learn lessons in a timely manner in practice if this involves embarrassing admissions of policy wrongs.

The Period of Debt Escalation: 1980–1993

In 1981 Greece became a fully fledged member of the EU, and this marked a wholly new period for economic and political developments in the country. Greece was one of the first non-founding countries to start connection talks with the European Common Market as early as 1961, but the process was abruptly suspended with the advent of the military dictatorship that lasted until 1974. Although membership of the EU was rightly viewed as an anchor of political and institutional stability for the newly restored democracy, it nonetheless fed and multiplied uncertainties over the capability of the economy to meet the challenge of integration.

The signs were all too visible: after a long period of systematic growth, Greece entered a period of recession in the late 1970s not only as a consequence of worldwide stagflation, but also because – on its way to integration with the common market – it had to dismantle its preferential system of subsidies, tariffs and state procurement by which several companies were kept profitable without being competitive. Soon after accession, many of these companies went out of business, and unemployment started to rise for the first time in many decades.

To face the slowdown, the government opted for a massive fiscal expansion that included demand-push policies to boost activity and the public underwriting of several ailing companies to maintain employment. The outcome was quite predictable: private debts turned into a chronic haemorrhage of budget deficits without any supply-side improvements. Similarly, the expansion of demand simply led to more imports and higher prices. Activity got stuck and Greece ended up in a typical stagflation, perhaps the quickest assimilation to European practices of the time.

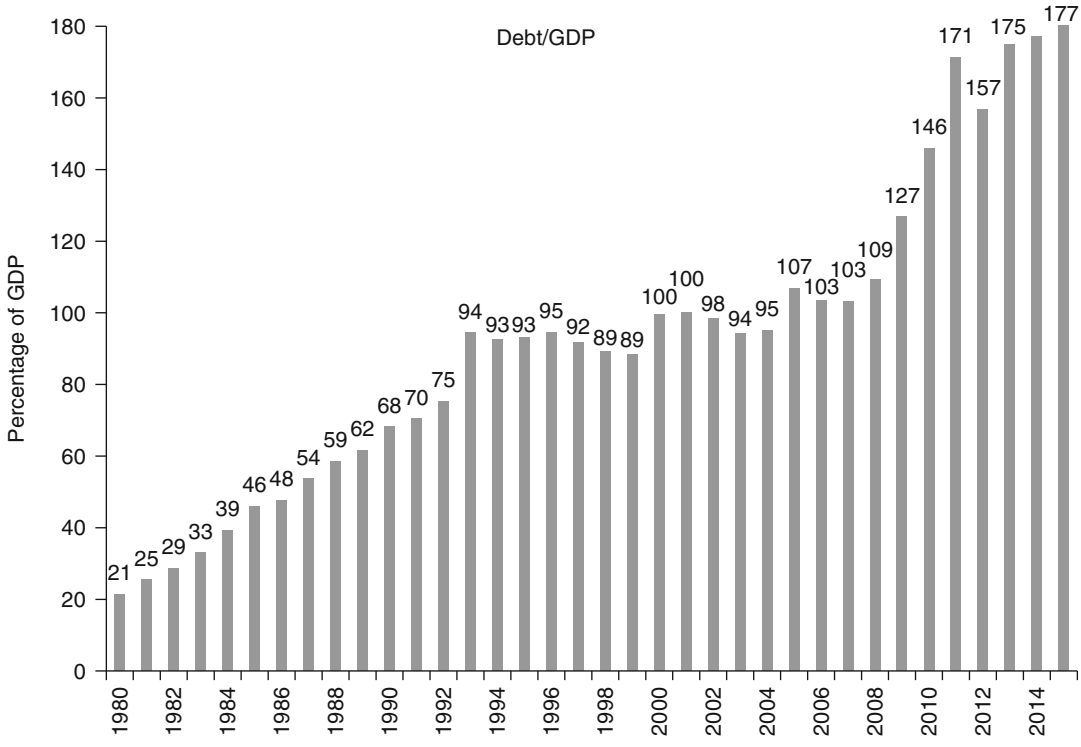
As a result, the arrival of Greece in the European Promised Land strangely coincided with the unleashing of a nightmare thought to be extinct after the Second World War: *rising public debt*.

Looking at Fig. 1, there are four distinguishable phases for the dynamics of debt: The first covers the period 1980–1993, during which public debt rose from slightly above 20% of GDP toward 100% in 1993. The second phase spans the period 1994–2005, in which public debt ends up again at around 100% of GDP after two mild reductions in between. The third phase covers the period 2006–2011, when public debt surpassed the 100% threshold, accelerated after 2008 and reached roughly 127% of GDP in 2009, triggering the crisis. At the beginning of 2012, Greek debt was partially cut through the process of Private Sector Involvement (PSI), and the hope was that it could be brought under control afterwards. But as none of the debt-augmenting factors was sufficiently dealt with, escalation continued after a brief (and small) respite in 2012. In the last phase 2012–2015, Greek debt approached 180% of GDP and the policy conundrum became more embarrassing, as the lack of growth has made even the return of the debt burden to the crisis-triggering levels of 2009 to look infeasible.

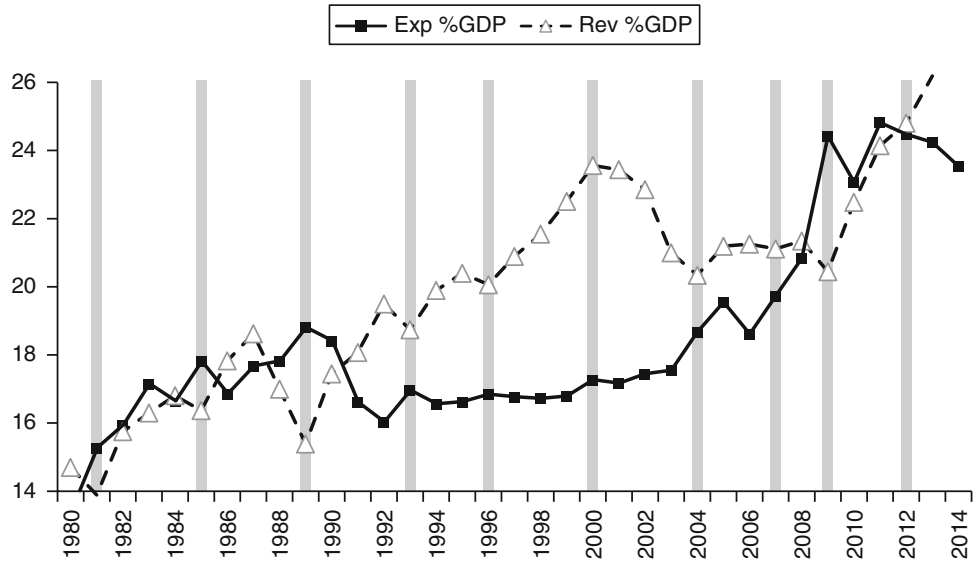
The above periodicity broadly coincides with substantial shifts in the context of economic policies, as suggested by developments in the fiscal patterns shown in Fig. 2 and in the current account depicted in Fig. 3 and briefly discussed below.

Regarding fiscal developments, the main characteristic of the first period was the substantial expansion of public spending and the concomitant rise in budget deficits and government debt. Revenues increased as a proportion of GDP, but were outpaced by the steadily growing expenditure. Both fiscal components appear to be volatile in the election years 1981, 1985 and 1989, suggesting the presence of a strong political cycle in public finances, as will be discussed later in more detail.

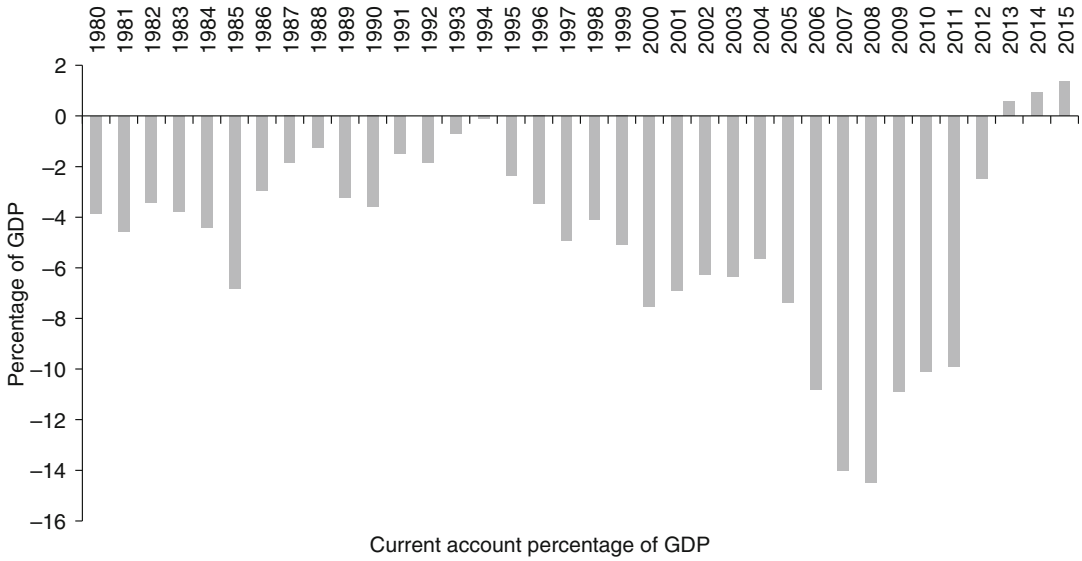
To maintain competitiveness, authorities had adopted, since the mid-1970s, a real exchange rate target with a crawling peg. After an automatic



Greek Crisis in Perspective: Causes, Illusions and Failures, Fig. 1 Greek public debt as percentage of GDP for the period 1980–2015 (Source: Ameco Eurostat 2015. General government consolidated gross debt: Excessive deficit procedure (based on ESA, 2010) and former definitions (linked series) (UDGGL))



Greek Crisis in Perspective: Causes, Illusions and Failures, Fig. 2 Primary public expenditure and total revenues as percentage of GDP in Greece, 1980–2014. Election years are denoted by bars. Dotted line denotes total public revenues (Source: AMECO Database, 2012, and Greek Budget Reports.)



Greek Crisis in Perspective: Causes, Illusions and Failures, Fig. 3 Current account in Greece as percentage of GDP, 1980–2015 (Source: IMF WEO Database, April 2015.)

wage indexation scheme was adopted in 1982, the only effect of the exchange rate policy was to fuel price increases and thus further aggravate trade deficits. To break the vicious cycle of depreciation and inflation, a discrete devaluation combined with a temporary wage freeze was implemented in 1983, but it was quickly superseded by a new phase of expansion as elections were approaching, driving public debt to even higher levels.

The external deficit approached 8% of GDP in 1985, an alarming threshold at that time, as witnessed by several Latin American economies

with similar imbalances that were serially collapsing. A coherent stabilisation programme was called for in October 1985, enforcing a discrete devaluation by 15%, a tough incomes policy and extensive cuts in public spending. The programme achieved a rise in revenues by beating several tax evasion practices and replacing previous indirect taxes with the more effective VAT system adopted by the EU. Public debt was stabilised, but only until the programme was finally abandoned in 1988 after being fiercely opposed from within the government and the ruling party.

The First Fiscal Crisis

Two general elections in 1989 failed to secure majority, thus leading to the formation of coalition governments, an event that was hailed as a confirmation of political maturing and an opportunity to overcome partisan differences on major issues. But self-indulgent admiration was short-lived, as stabilisation policies are notoriously difficult to implement through party coalitions because each party tries to avoid the cost falling on its own constituency. Greece was no exception to the rule, and the economy suffered a major setback in 1989, far more serious than previous fiscal failures.

Two episodes are characteristic of how rhetoric designed to please everybody in combination with naïve policies can lead to disaster: despite looming deficits, in 1989 the coalition government decided to abolish prison terms for major tax arrears, hoping to induce offenders to repent and reconsider their strategy. Expectedly, the move was interpreted the other way around, as a signal of relaxed monitoring in the future, thus encouraging further evasion.

Another bizarre policy was to cut import duties for car purchases by repatriates returning to Greece after the collapse of the Soviet Union. The measure was viewed as a gesture to facilitate mobility back in the motherland, but it was quickly turned into a black market scheme. For a small bribe, immigrants were purchasing luxury cars only to immediately resell them to rich clients who could thereby avoid the duty tax. The budget was deprived of badly needed revenues and evaders had yet another reason for celebration.

As a result, revenues collapsed and the country suffered a major fiscal crisis, until a majority government was elected in 1990 that enacted a new stabilisation programme. Despite substantial cuts in spending and a rise in revenues, public debt as a ratio to GDP continued to rise because of the higher cost of borrowing worldwide and a stagnant output. The sharp rise in 1993 in particular was due to the inclusion of extensive debts initially contracted by public companies under state guarantees but finally underwritten by the budget. Except for the electoral years 1989–90, fiscal consolidation significantly improved the current account, and such a rarity as a balanced external position was reached in 1994.

Debt Stabilisation and EMU Membership

Although Greece was a signatory of the Maastricht Treaty in 1991, it was far from obvious whether, how and when the country could comply with the nominal convergence criteria required to join the Economic and Monetary Union. Public deficits and inflation were galloping at two-digit levels and there was great uncertainty about the viability of the exchange rate system; a lengthy fiscal analysis of the period is provided by Christodoulakis (1994).

In May 1994, capital controls were lifted in compliance with European guidelines and this prompted fierce speculation in the forex market. Interest rates reached particularly high levels and the Central Bank of Greece exhausted most of its reserves to stave off the attack; for an account of the successful defence see Flood and Kramer (1996). This episode proved to be a turning point for the determination of Greece to pursue accession to EMU in order to be shielded by the common currency and avoid similar attacks in the future. Soon afterwards the ‘Convergence Programme’ was adopted that set time limits to satisfy the Maastricht criteria and included a battery of reforms in the banking and the public sectors.

International markets were not impressed and continued to be unconvinced about exchange rate viability. With the advent of the Asian crisis in 1997 spreads rose again dramatically and – after months of credit shortages – Greece finally decided to devalue by 12.5% in March 1998 and subsequently entered the Exchange Rate Mechanism, wherein it had to stay for two years. Greece still had a lot of housekeeping to do, and – unable to join the first round of eurozone countries in 1998 – was granted a transition period to comply with the convergence criteria by the end of 1999.

After depreciation, credibility was further enhanced by structural reforms and reduced state borrowing, so that when the Russian crisis erupted in August 1998, the currency came under very little pressure. Public expenditure was kept below the peaks it had reached in the previous decade and was increasingly outpaced by the rising revenues and various one-off receipts. Tax collection was enhanced by the introduction of a scheme of minimum turnover on SMEs, the elimination of a vast number of tax allowances, the imposition of a new levy on large property and a re-reorganisation of the auditing system. Proceeds were further augmented by privatisation of public companies and, as a result, public debt fell to 93% of GDP in 1999. Although still higher than the 60% threshold required by the European Treaty, Greece benefited from the convenient interpretation that it suffices '*to lean toward that level*', as previously used by other countries – such as Italy and Belgium – in their own way to enter EMU.

The Implementation of Market Reforms

In the 1980s, structural reforms were hardly on the agenda of Greek economic policy. In fact, for most of the period the term was a misnomer used to describe further state intervention in economic activity, rather than market-oriented policies as practised in other European countries. Market reforms were introduced for the first time in 1986, aiming at the modernisation of the outmoded banking and financial system in compliance with European directives. A major reform in social security took place for the first time in 1992, curbing early retirement and excessively generous terms on the pension/income ratios.

Throughout the 1990s, various reform programmes were aimed at the restructuring of public companies whose chronic deficits had contributed to the fiscal crisis in 1989. Privatisation was attempted through direct sales of state-owned utilities as the quick way to reduce deficits. Despite some initial success, the programme was fiercely opposed by the trade unions of public companies and eventually led to the demise of the government. Privatisations were conveniently branded as sell-outs, and it took a few more years for the concept to reappear on the political agenda.

A new wave of reforms was launched after 1996 in the course of the 'Convergence Programme'. State banks were privatised or merged, dozens of outmoded organisations were closed down and a series of IPOs – taking advantage of the stock market bonanza – provided capital and restructuring finance to several public utilities. Other structural changes included the lifting of closed-shop practices in shipping, the entry of more players into the mobile telephony market and a series of efforts to make the economic environment more conducive to entrepreneurship and employment.

Post-EMU Fatigue

After 2000, Greece emulated some other euro area members in exhibiting a 'post-EMU fatigue' and the reform process gradually slowed down. As shown in Fig. 9 (see later), proceeds from privatisation peaked in 1999, but subsequently remained low as a result of the contraction in capital markets after the dot.com bubble and the global recession in 2003; for an extensive discussion of reforms in Greece over the period 1990–2008 see Christodoulakis (2012).

An attempt in 2001 to deeply reform the pension system led to serious social confrontations and was finally abandoned. Although replaced by a watered-down version one year later, the failure left a mark of reform timidity for many years. Two other mild reforms followed in 2006 and 2010, but

the social security system is still characterised by inequalities, inefficiencies and structural deficits that exert a substantial burden on the general government finances.

The fatigue spread more widely after the Olympic Games in 2004. With the exception of the sale of Greek Telecom to the German state company and the privatisation of the national air carrier after a decade of failed attempts, most other reforms consisted of small IPOs with no structural spillovers to the rest of the economy.

Why Debt Reduction was Insufficient

Despite having achieved substantial primary surpluses throughout 1994–2002 – and around 1999 in particular – public debt over the same period fell only slightly. There are three reasons to explain this outcome.

First, during this period the government had to issue bonds to accumulate a sufficient stock of assets for the Bank of Greece as a prerequisite for its inclusion in the euro system, and this capital injection led to a substantial increase in public debt without affecting the deficit.

Second, after a military stand-off in the Aegean in 1996, Greece increased defence procurement to well above 4% of GDP per year. In line with Eurostat rules, the burden was fully recorded in the debt statistics at the time of ordering, but only gradually in the current expenditure following the pattern of actual delivery of equipment. This practice created a considerable lag in the debt-deficit adjustment, and in 2004 the government enforced a massive revision of the deficit figures by retroactively augmenting public spending on the date of ordering, prompting a major dispute over the quality and integrity of the statistics of public finances in Greece. Although a decision by Eurostat in 2006 made the delivery-based rule obligatory for all countries, Greece did not withdraw the self-inflicted revision. As a consequence, deficits were statistically augmented for 2000–2004 and scaled back for 2005–2006 relative to what they should have been otherwise, in an awkward demonstration of political interference.

The third reason was the strong appreciation of the yen/euro exchange rate by more than 50% between 1999 and 2001. This significantly augmented Greek public debt as a proportion of output due to the fact that substantial loans were contracted in the Japanese currency during the 1994 crisis. To alleviate this exogenous deterioration, Greece entered a large currency swap in 2001 by which the debt to GDP ratio was reduced by 1.7% in exchange for a rise in deficits by 0.15% of GDP in subsequent years, so that the overall fiscal position remained unchanged in present value terms. Although the transaction had no bearing on the statistics for 1999 on which EMU entry was assessed, Greece suffered extensively from criticisms that mistook the swap as a ploy to circumvent a proper evaluation. The values shown in Fig. 1 are net of swap effects, and this partly explains the peak in 2001.

The Chronic Current Account

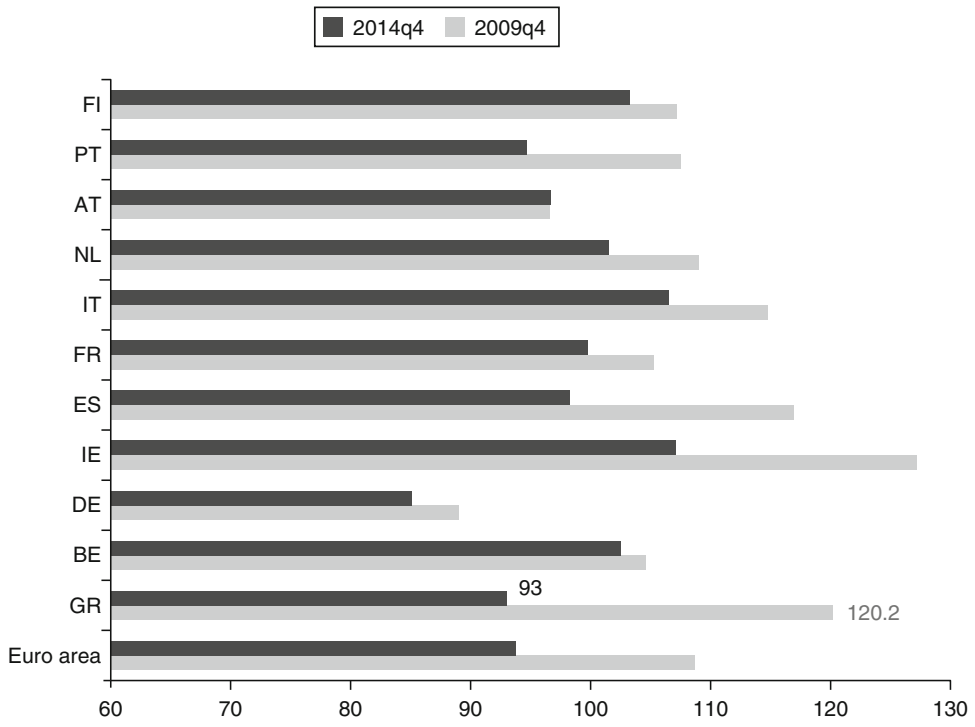
After the eurozone became operational, hardly any attention was paid to current account imbalances, regarding Greece or any other deficit country. Even after they reached huge proportions, external disparities in the euro area continued to remain surprisingly unnoticed from a policy point of view. It was only in the aftermath of the 2008 crisis that policy bodies in the EU started emphasising the adverse effects that external imbalances may have on the sustainability of the common currency (see for example European Commission 2009).

The reason for this complacency was not merely that devaluations were ruled out by the common currency. A widespread – and unwisely comfortable – view held that external imbalances were mostly demand-driven effects and, as such, they would sooner or later dissipate as a result of ongoing fiscal adjustment in member states. When, for example, Blanchard and Giavazzi (2002) asked whether countries such as Portugal or Greece should worry about and take measures to reduce their current account deficits they '… conclude(d), to a first order, that they should not'. A few years later this proved to be just another misguided assessment;

Blanchard (2006) – overturning his previous optimism – remarked that current account deficits were steadily increasing within the euro area and urged immediate action, otherwise '. . .implications can be bad'. And indeed they were.

Although it improved for a while after the country joined the common currency, the subsequent vast deterioration in the Greek current account played a crucial role in inviting the global crisis home. The reason behind the initial containment was that factor income flows from abroad increased as a result of extensive Greek foreign direct investment in neighbouring countries while labour immigration kept domestic wage increases at bay. The deficit started to deteriorate after 2004 as domestic demand peaked in the post-Olympics euphoria, inflation differentials with other eurozone countries widened and the euro was further appreciated. Unit labour costs increased and as shown in Fig. 4 the relevant index rose by 20.2% in the period 1999q1–2009q4. As *ex post* wisdom, it is worth noticing from the same graph that an erosion of competitiveness took place in *all* other eurozone countries that later sought bail-out agreements (Ireland by 27% and Portugal 8%) or were considered to be at the risk of seeking one (Spain by 17% and Italy by 15%).

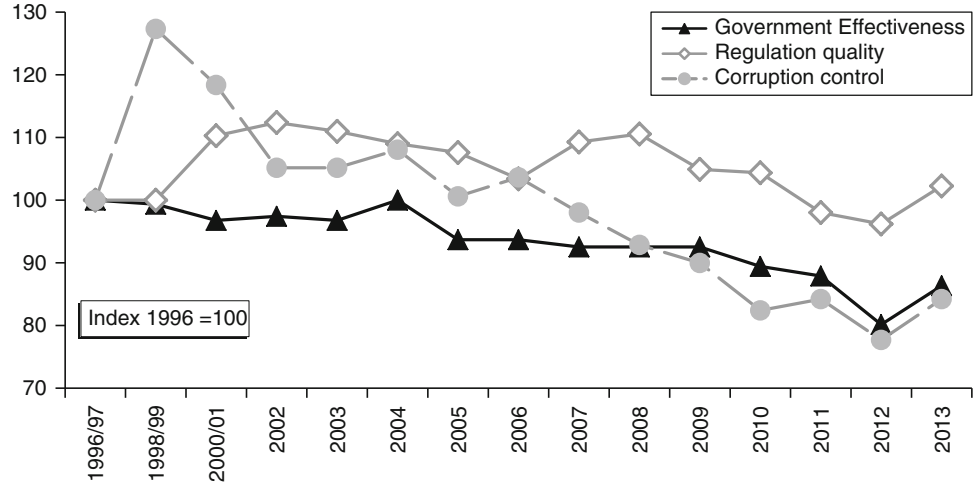
Compared to Germany, the Greek unit labour cost increased by 32% relative to the 1999 level, thus causing significant bilateral imbalances. However, this erosion was gradual and cannot have been the single reason for the rapid deterioration experienced after 2006. Other factors affecting the investment environment, such as the quality of the regulatory framework, elimination of corruption practices and overall government effectiveness might also have been crucial in shaping productivity and competitiveness. Using the Worldwide Governance Indicators published by the World Bank as proxies for how the above



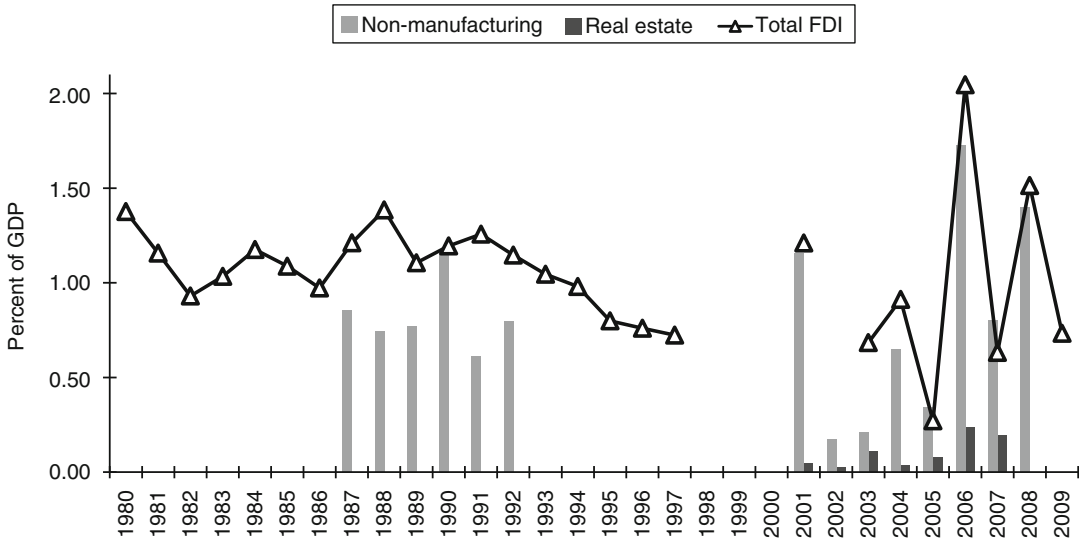
Greek Crisis in Perspective: Causes, Illusions and Failures, Fig. 4 Developments in unit labour costs in the eurozone (1999q1 = 100) (Source: ECB, Harmonised competitiveness indicators based on unit labour costs indices for the total economy, for 2009q4 and 2014q4. The effect of Greek ULC on competitiveness is mainly due to the wage-cuts implemented through 2011–2012.)

factors evolved during the period from 1996 to 2013, Fig. 5 shows that, despite some improvement in the first years of EMU, there was a noticeable decline thereafter.

These developments were pivotal to the poor performance of Greece in attracting foreign direct investment in spite of the substantial fall in interest rates and the facilitation of capital flows within the eurozone. As depicted in Fig. 6, FDI expressed as percent of GDP hardly improves during the last decade relative to the 1980s. The composition has also changed, as most of the FDI inflows were



Greek Crisis in Perspective: Causes, Illusions and Failures, Fig. 5 Quality indicators affecting the economic climate 1996–2013 (*Notes*: Indicators are measured in various units with higher values corresponding to better outcomes; to ease comparison all are indexed at 100 in 1996. Source: World Bank, WGI various editions. Figures for 1997, 1999 and 2001 are not available.)



Greek Crisis in Perspective: Causes, Illusions and Failures, Fig. 6 FDI inflows to Greece expressed as percentage of GDP (*Note*: Missing observations are due to non-availability and do not necessarily imply poor flows. Source: OECD, FDI statistics. After the implementation of the austerity programme, FDI flows virtually ceased due to the increasing uncertainty.)

directed to non-manufacturing sectors and, pointedly, with an increasing allocation to real estate that further aggravated the strain in the current account.

It is a well-established fact that when new investments are directed mainly to the tradeable sectors this leads to substantial productivity improvements and favours net exports. In contrast, investments going mostly into the real estate sector boost aggregate demand, raise prices, cause the real exchange rate to appreciate and hinder competitiveness. These developments manifest a major failure of Greece – and for that matter of other eurozone countries – to exploit the post-EMU capital flows in order to upgrade and expand production; for details see a study by Christodoulakis and Sarantides (2011) who use the differentiation in composition and the asymmetry in the destination of FDI to explain the diverging patterns of external balances in the eurozone countries before the 2008 crisis.

Unprepared for the 2008 Crisis

The fiscal decline started with the disappearance of primary surpluses after 2003 and culminated with rocketing public expenditure and the collapse of revenues in 2009, as shown in Fig. 2. Revenues declined as a result of a major cut in the corporate tax rate from 35% to 25% in 2005 and extensive inattention to the collection of revenues.

Such decisions were making it increasingly evident that stabilising the economy was not a policy priority of the government, and further actions soon confirmed the assumption: concerned over the rising deficits in 2007, it sought a fresh mandate to redress public finances but – despite securing a clear victory – no such action whatsoever was taken after the election. Only a few months before the global crisis actually erupted, the government claimed that the Greek economy was 'sufficiently shielded' and would stay immune to the reverberations of international shocks. Even after September 2008, the Government was for a long time ambivalent as to whether to implement a harsh programme to stem fiscal deterioration or to expand public spending to fight off the prospect of recession. A final compromise at the end of the year included a consumption stimulus combined with a bank rescue plan of h5 bn and a pledge to raise extra revenues. The first two were quickly implemented, whilst the latter was forgotten.

Weakened by internal divisions, the government continued to be indecisive on what exactly to do and, after a defeat in the European elections in June 2009 it opted for yet another general election in October 2009, asking for a fresh mandate to address the mounting economic problems. In practice, the election period turned to be an opportunity for further largesse rather than of preparation on how to contain it. The fiscal consequences were stunning: total public expenditure was pumped up by more than 5 percentage points, exceeding 31% of GDP at the end of 2009. (In levels, it exceeded h62 bn, i.e. twice the size in 2003). The rise was entirely due to consumption, as public investment remained the same at 4.1% of GDP; details on how public spending was ballooned are given in Christodoulakis (2010).

Total receipts in 2009 collapsed by another 4% of GDP as a result of widespread neglect in collection and the fact that privatisation proceeds turned negative since the government had to finance the emergency capitalisation of Greek banks. The deficit of general government spiralled and its figure was serially revised from an estimated 6.7% of GDP before the elections to 12.4% in October 2009, and finally widening to 15.4% of GDP by the end of the year. It was only then that European authorities stopped their wait-and-see attitude and issued a number of warnings against the spending.

Post-election Inaction

In spite of the gathering storm in the autumn of 2009, the newly elected Government was far from being determined to achieve immediate fiscal consolidation, constrained as it was by its pre-electoral rhetoric that 'money exists' and its ideological aversion to controlling trade union demands in public enterprises. Trapped in such

unrealistic mentalities, the December budget for 2010 surprised everybody by including an *expansion* of public expenditure and completely *excluding* privatisations, rather than the other way around. Seeing that no appropriate action had been taken to deal with the situation, rating agencies downgraded the economy. This sparked massive credit default swaps in international markets and the crisis loomed.

The problem Greece faced at that time was an acute shortage of financing for the deficit, not yet one of debt sustainability as it later turned out to be. In this regard, a significant opportunity to defuse the crisis was missed by the government and European authorities alike. In order to reduce the risk of spillovers to other markets after the credit crunch in 2008, the ECB had invited private banks of euro member states to obtain low-cost liquidity by using sovereign bonds from their asset portfolio as collateral securitisation; see De Grauwe (2010) for a positive assessment of this policy. As a result of this credit facilitation, yields on Treasury Bills remained exceptionally low. But instead of borrowing cheaply in the short term as a means of gaining time to redress the fiscal situation, the government kept on issuing long maturities despite the escalation of costs. This had dramatic consequences on the perception of the crisis by international markets. Commenting on the cost of confusion, Feldstein (2012) aptly notes that:

> "What started as a concern about a Greek *liquidity problem* – in other words, about the ability of Greece to have the cash to meet its next interest payments – became *a solvency problem*, a fear that Greece would never be able to repay its existing and accumulating debt", (my emphases).

Adding injury to misjudgement, the situation was further undermined when the ECB threatened to refuse collateral status for downgraded Greek bonds, hence fuelling fears that domestic liquidity would shrink and precipitating a capital flight from Greek banks. Three months later the rating requirement was dropped for all eurozone countries, but the damage was no longer reversible. In early 2010, borrowing costs started to increase for both short- and long-term maturities, Greece had become a front-page story worldwide and the countdown began. Despite the belated ECB generosity, the government was financially exhausted and in April 2010 sought a bailout.

Missing Defence and the Role of External Deficits

The global financial crisis in 2008 revealed that countries with sizeable current account deficits are vulnerable to international market pressures because they risk having a 'sudden stoppage' of liquidity. Recent studies show that highly indebted EMU countries with large external deficits are found to experience the highest sovereign bond yield spreads. Along this line, Krugman (2011) recently suggested that the crisis in the southern eurozone countries had rather little to do with fiscal imbalances and rather more to do with the sudden shortage of capital inflows required to finance their huge external deficits.

This explains why immediately after the crisis sovereign spreads peaked, mainly in economies with large external imbalances, such as Ireland, Spain, Portugal and the Baltic countries, which were under little or no pressure from fiscal deficits; for a discussion of the effects of the credit crunch in emerging markets with large current account deficits, see Shelburne (2008). In contrast, countries with substantially higher debt burdens, but without external imbalances, such as Belgium and Italy, experienced only a small increase in their borrowing costs at that time.

Greece happened to have a dismal record on both deficits and its exposure to the international credit stoppage was soon transplanted into a debt crisis. The current account went into free-fall after 2006 when three factors intensified: domestic credit expansion accelerated and disposable incomes were enhanced by the tax cuts, while capital inflows from the Greek shipping sector peaked as a result of the global glut and the huge rise in Chinese freight. The external deficit exceeded 14% of GDP in 2007 and 2008 and still no warning was voiced by any authority, domestic or European. In fact, quite the opposite happened: responding to pleas of car dealers, the Greek government decided to reduce surcharges

on imported vehicles in an attempt to revive the market, while other fellow governments – at least those from car-making countries – failed to notice the pro-cyclical character of the measure. Replicating history back in 1989, the unfortunate act to facilitate car purchases in order to favour particular groups again caused a significant deterioration of both the external and the public deficits. Additionally, nobody missed the signalling about the true priorities of the Government and the pre-electoral spree followed as described above.

The events that took place in Greece – and to a lesser extent in other euro area countries a few months later – revealed a wide gap of policy ineffectiveness and unpreparedness by the European authorities as well. The inability of the eurozone authorities to understand the causes and foresee the dynamics of external imbalances among its member states was only paired by the lack of institutions to face their consequences in case they become too dramatic. After the event, one could only hypothetically assess the degree of damage control that the operation of newly founded institutions could have allowed: for example, the Fiscal Compact would have warned European authorities about the explosive Greek finances after 2007 as described before. The implementation of the Banking Union might possibly have detected the overexposure of Cypriot banks and contained the panic after their crash.

But such thoughts rarely disturbed the anti-inflation single-mindedness prevailing at that time in Frankfurt or Brussels headquarters. In the jubilant pre-crisis atmosphere of capital abundance worldwide, European authorities never considered the possibility of a sudden stop that could accelerate into a credit crunch. In fact, the only fear expressed in policy circles was about the opposite: that credit was so much oversupplied as to threaten the low inflation target, and mechanically this lead the ECB to raise the interest rate in July 2008 to the fearsome level of 4.25%. When two months later the global crisis erupted full steam, this controversial move simply made the credit pressure higher for the current account deficit countries and – after an exhaustive defence with ever rising borrowing costs – one after the other started seeking bailout agreements. Greece was the first to enter the chorus, followed by Portugal, Ireland and Spain – though the latter with fewer austerity requirements imposed on the others.

Two Important Policy Facts

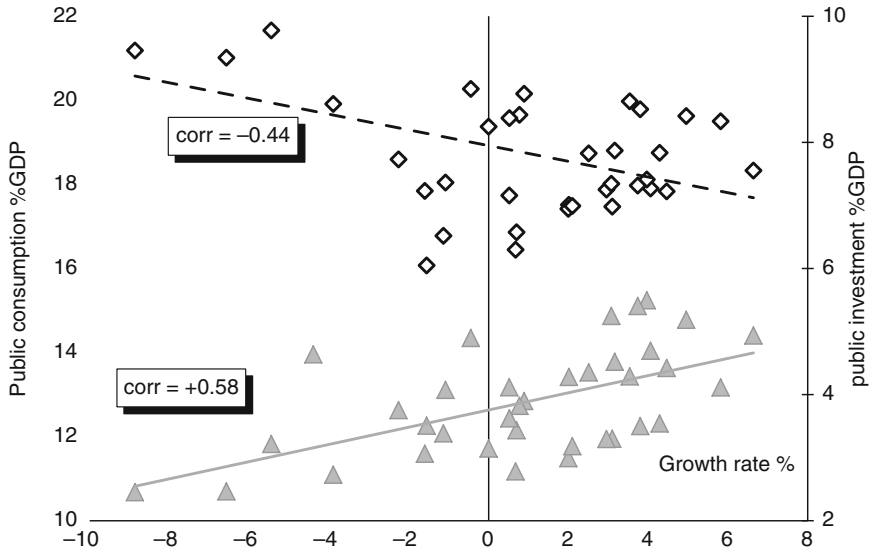
Two essential facts emerge from the historical account of fiscal developments in Greece. One is the fact that in periods of recession counter-cyclical activity usually takes the form of increased consumption, not public investment, and this has detrimental effects on public and external deficits without contributing to higher growth. Another recurring characteristic is the propensity of governments to increase public spending and to tolerate lower revenues in election years.

Cyclic Nature of Public Spending

As an indication of how the two main components of government spending behave over the economic cycle, public consumption and public investment expressed as proportions of GDP are juxtaposed with the growth rate; see Fig. 7. Public consumption is found to have a strong negative correlation with growth rates, suggesting a counter-cyclical pattern. This finding implies that periods of economic downturn are likely to be associated with higher public consumption due to increased benefits and programmes to contain unemployment. In a situation of fixed public employment and nominal wage resistance, public consumption is expected to rise further relative to GDP.

On the other hand, public investment shows a strong positive correlation with the growth rate. This implies that in a downturn public investment is likely to fall, thus hindering the resumption of growth and spreading more recession in the economy.

A clear manifestation of such behaviour over the cycle took place in the recent post-crisis years. With recession deepening year after year, the government, rather than curtailing the public sector,



Greek Crisis in Perspective: Causes, Illusions and Failures, Fig. 7 Growth rate correlations with public consumption (LHS) and public investment (RHS) expressed as percentage of GDP. (Source: Greek Government Budget Reports, various editions.)

found it more expedient to cut public investment in order to control the deficit. No wonder then that recession was made much worse.

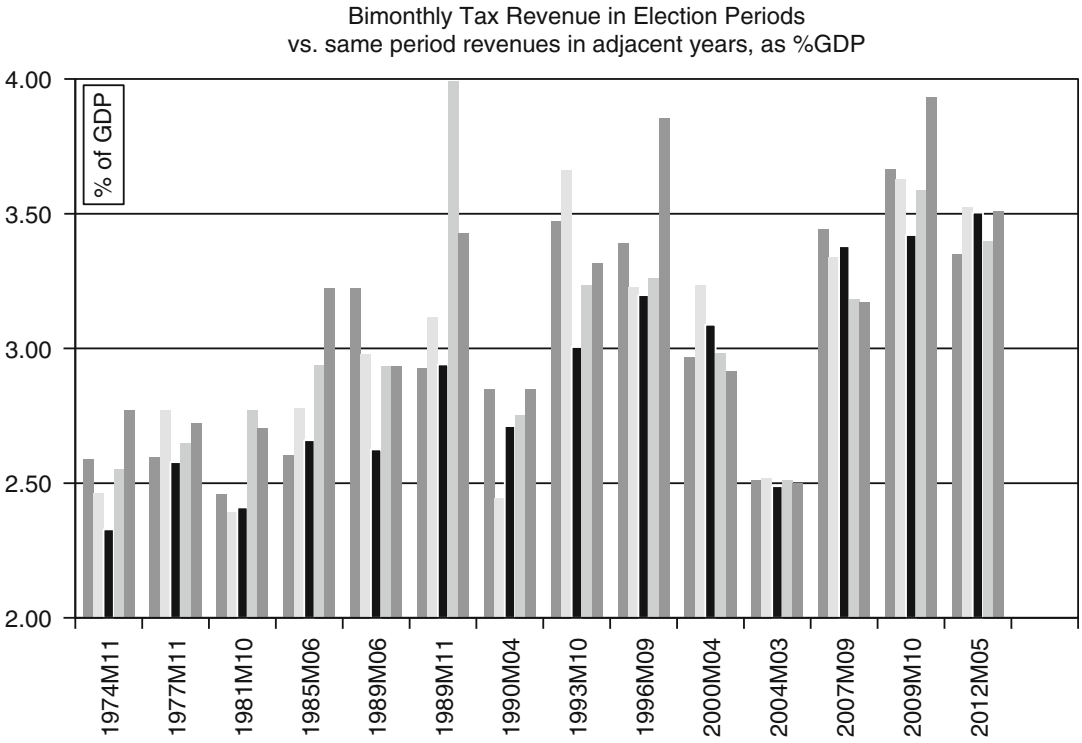
Electoral Cycles

The Greek economy was often subject to the electoral cycle, as before the elections most incumbent governments tried to appeal to voters by a variety of opportunistic policies, thus inflicting non-trivial (and usually permanent) fiscal losses. Practices included extra appointments of party affiliates, grants to favourable groups and allocation of petty projects to local constituencies, all of which affect current or future expenditure.

It can readily be seen from Fig. 2 that spending rose during election years in the 1980s. As deficits widened, the economy had to enter a period of stabilisation that was usually terminated just before the next election and routinely accompanied by Finance minister's replacement. During the debt escalation in 1980–93 there were four stabilisation programmes and ten Finance Ministers – usually one to pursue the programme and then a successor to denounce it and prepare for the next period of spending tips. Although the electoral cycle subsided in the period before and after EMU membership, it returned full-steam in the elections of 2009. In the elections of 2012 the effect was dissipated.

Apart from direct actions on the expenditure side, the empirical evidence suggests that slacker tax auditing around elections causes further fiscal deterioration. An extensive investigation by Skouras and Christodoulakis (2014) found that flaws in tax collection arise either as a result of deliberate relaxation of firm-level auditing as a signal to political supporters or as an indirect consequence of the slackness prevailing in public administration around elections.

Considering that a typical pre-election period in Greece has a duration of around 40 days, Fig. 8 compares the revenue in the two months of the election period in each electoral year with the same two months in adjacent years. Simple inspection shows that in most of the elections held between 1974 and 2009, average bimonthly revenues expressed as percentage of GDP were lower than the average of the respective figures in the two adjacent years (with only two slight exceptions: in 2000, coinciding with the entry to



Greek Crisis in Perspective: Causes, Illusions and Failures, Fig. 8 Comparison of bimonthly tax revues in pre-election periods. Revenues are calculated for the period of two months before each election as percentage of annual GDP. Each election year (N) is shown in black and compared with revenues collected over the same period during the previous ($N - 2$, $N - 1$) and the following ($N + 1$, $c + 2$) years shown in grey. Frequency is bimonthly to account for the fact that the pre-election period lasts for 30–40 days; thus it extends over the prior as well as the poll month. Data are not seasonally adjusted, thus they reflect within-year variations (Source: Skouras and Christodoulakis (2014), where further details are available.)

EMU, and 2007, because it is compared with another – and a lot worse – electoral period in 2009). In the same study it is estimated that pre-electoral misgovernance causes a loss in revenues estimated to be 0.18% of GDP in each election year. For the 13 elections that have taken place in the period 1974–2009, this amounts to more than h5 bn at 2010 prices.

An *ex post* Assessment of the memorandum

EU authorities seemed to be unprepared to react promptly and concertedly to the Greek problem and undertook action only when they recognised the risks it posed for the banking systems of other European states. After difficult negotiations, a joint loan of h110 bn was finally agreed in May 2010 by the EU and the IMF to substitute for inaccessible market borrowing. The condition was that Greece follow a Memorandum of fiscal adjustments to stabilise the deficit and structural reforms to restore competitiveness. More details are given in the Appendix. In the event of success, Greece would be ready to tap markets in 2012 and then follow a path of lowering deficits and higher growth. But year after year the actual outcome remained far below from the projections and the economy contracted fiercely. An explanation is attempted below.

The Failure in Fiscal Adjustment

The decline of revenues as a share of GDP after 2007 and the collapse of the collection mechanism

in 2009 in particular were instrumental in the explosion of public debt and deficit thereafter. Strangely enough, no serious effort was undertaken to remedy the situation after the elections. The ministerial post in the Inland Revenue remained empty for more than a year and two top executives resigned in protest that their proposals to beat tax evasion were turned down. The government opted for an increase in the VAT rate from 19% to 23% in spring 2010 and, as a result, CPI inflation jumped to 4.5%, further cutting purchasing power amid recession. The only result was that activity was reduced and revenues did not rise.

The government continued to act in a positive feedback loop, with lower revenues prompting higher taxation, with this in turn causing further evasion. Unable to raise efficiency and under pressure to collect revenues, it imposed a heavy increase in fuel tax, substantial consumption surcharges and finally a lump-sum tax in exchange for settling previous arrears. Once again, tax revenues ended up far below the target in a typical manifestation of elementary Laffer curve predictions.

Only by the end of 2011 was it recognised that further tax measures were no longer viable and that attention should shift to collection efficiency. In its assessment of progress, the European Commission task force warned that

> "... tax and expenditure measures substantially compress the households' disposable income and significantly tighten their liquidity constraints" (European Commission 2011, p. 2).

But that was no more than an empty warning, because at the same time the government was forced by the very same task force to retroactively raise the tax rate on the self-employed and impose a new levy on property in order to make up for falling revenues.

Regarding public expenditure, a more optimistic picture emerged, but at a huge cost in terms of growth and efficiency. Soon after the elections, the government made clear signals that it had no real intention of containing the oversized public sector. Numerous appointments that were made before elections through a highly disputed process were nevertheless approved by the new incumbent, and a widely publicised operation to abolish and merge outdated public entities has made no real progress to date. A novel scheme to push older staff onto standby status with a fraction of their salary misfired, as it was soon discovered that most on the list were exploiting the incentives of the system for an early retirement. After the fiasco, the government announced a lengthy process of evaluation in the public sector as a precondition for staff redundancies, but this again produced no concrete results. It was only in 2013 that some serious measures started to be taken, and a number of layoffs were implemented in the public sector. But this was too little and too late: public sector unions organised a series of counter-actions both in the streets and in the courtrooms and several layoffs were overturned. In any case, the new government that emerged from the 2015 elections started reinstating the dismissed personnel, triggering a new cycle of political confrontation and fiscal burden.

In the absence of any structural adjustment in the public sector, the reduction of spending was achieved by imposing universal cuts in salaries, and this led to widespread shirking practices. Another unusual tool for keeping expenditure low was to cut the budgetary co-financing of the European Community Support Framework, thus reducing public investment at a time when it was most needed to induce some growth in the economy. After the decision by the European summit in July 2011, Greece was freed from the co-financing obligation, but when the new practice started to be implemented at the end of 2011 it was already too late to rectify the damage done to economic activity.

The Limits of Structural Adjustment

In order to rebalance the economy onto a more competitive path, the Memorandum agreement envisaged a long list of structural reforms, ranging from reforming the social security system to removing closed-shop vocational practices, and from cutting red tape to liberalising the licensing process for lorry and taxi drivers. The pension

reforms initially succeeded in harnessing the deficits in the social security funds, but they soon reappeared when a wave of retirement took place in anticipation of imposing further age extensions in the future.

Most of the reforms were either abandoned or backfired. For example, the opening-up of lorry licences failed to reduce transportation costs and enhance competitiveness in practice, despite the severity of clashes with trade union hardliners. The reason for this was that insiders took advantage of a two-year postponement and decided to maximise rent-seeking by withdrawing previous price concessions. Besides, the economic gloom was thwarting potential investors by making the upfront cost of setting up a new business too high.

A similar attempt to open-up the taxi licensing system was abandoned after a protracted clash with insiders in the summer 2011 that seriously damaged tourism in its peak period. In other professions, such as law and pharmacy, there was only a token liberalisation without any reduction in consumer prices. Recognising this failure, the new conditionality programme imposed a regressive mechanism with the aim of reducing the overall profit margin to below 15% (see Memorandum II 2012, para 2.8, 'Pricing of medicines').

Seeing that the structural adjustment programme was derailed, the Memorandum sought alternatives. To enhance competitiveness in the labour market, liberalisation measures extended part-time employment, imposed wage cuts across the board and removed collective bargaining agreements. Despite lowering labour costs by 12%, enterprises were overwhelmed by recession and unemployment became rampant, exceeding 17% of the labour force by the end of 2011. As with the positive feedback mechanism on the tax front, the rise in unemployment invited a new round of wage cuts in the private sector, further shrinking disposable income and fuelling new waves of social protest. Although the IMF mission in the autumn 2011 was explicitly admitting that 'accelerated private sector adjustment... would likely lead to a downward spiral of fiscal austerity, falling incomes and depressed sentiment', it nevertheless urged further such measures in order to achieve a '... critical threshold of reforms needed to transform the investment climate' (IMF 2011). Instilling some growth to the real economy so as to kick-start activity and facilitate the implementation of market reforms was not a top priority for the programme overseers, even after admitting the previous failures.

The Failure of Privatisations

The failure of the privatisation programme is worth commenting on, as it reveals an unusual combination of strong rhetoric in theory with complete apathy in practice. Immediately after the elections in 2009, the government showed that it had no intention of curbing the wider public sector. Its lack of resolve to tackle the excessive demands of public trade unions was made manifest in a dispute with a newly arrived investor in the Piraeus Port Company. The government succumbed to paying enormous compensation for early retirement as a condition that the investment went ahead. No privatisation target was included in the 2010 budget and none was actually implemented.

Thus it was viewed as a major shift of policy when the government agreed in March 2011 to adopt a large-scale privatisation plan of h50 bn during the period 2011–2015, or roughly 4% of GDP per annum. The plan included extensive sales of public real estate, privatisations of public enterprises in the energy sector and private partnerships in the operation of airports and ports throughout Greece.

After months of procrastination a market-friendly privatisation fund was finally set up to replace the ineffectual authority that was in charge before, but its determination was this time hindered by adverse market conditions. With asset prices falling to abysmal levels, privatisations would be probably embarrassing in political terms and inadequate in terms of revenue, but in practice there was no real demand, as capital flight continued to be fuelled by fears of abandoning the eurozone and funds from abroad were not coming for the same reason. Despite initial ambitions, the programme achieved only a fraction of its target in 2011. By selling an option on Greek Telecom and

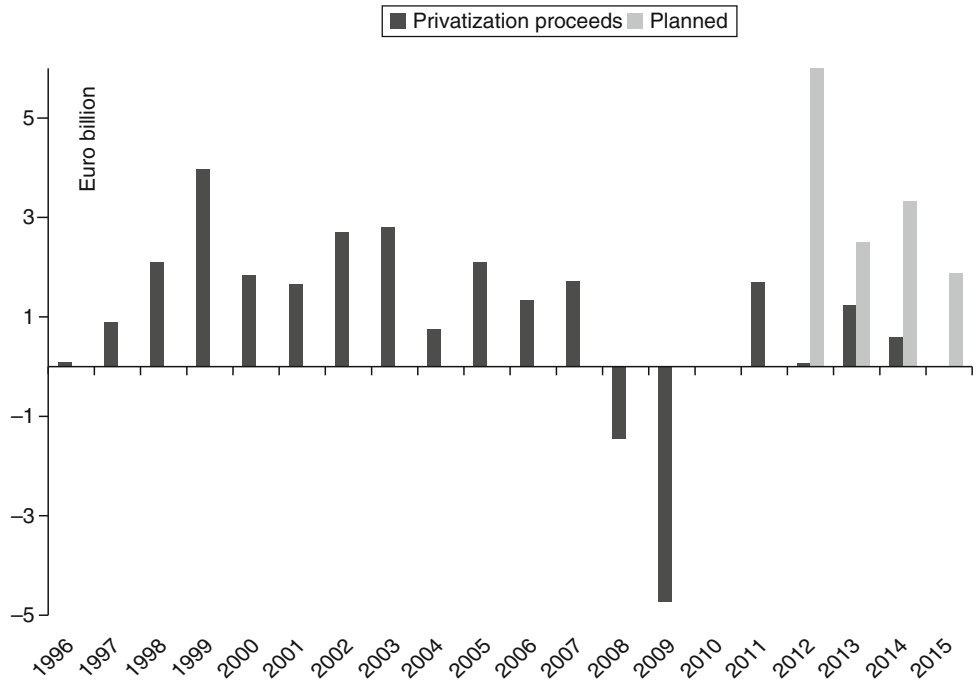
future rights to the National Lottery an amount of h1.7 billion was collected.

In 2012 the Government abandoned these infeasible targets and the programme was downscaled to a meagre h2.8 bn, just over a quarter of the amount initially announced. Alas, the outcome was again disappointing, as proceeds collected in that year reached only h69 million; i.e. only 2.4% of the envisaged target. A new institution was set up invested with the portfolio of all public companies shares and public real estate, and tasked with accelerating the privatisation process. The Hellenic Republic Asset Development Fund (TAIPED) fared better than before, but nevertheless was still some way from making privatisations a mechanism for attracting major investment initiatives. As shown in Fig. 9, misfiring continued, and in 2015 the whole programme was stalled by the new government.

The Second memorandum Conditionality and Ways out of the Crisis

Faced with a deepening recession and a failure to produce fiscal surpluses sufficient to guarantee the sustainability of Greek debt, the EU intervened twice to revise the terms of the Memorandum. In the first major intervention in July 2011, the amount of aid was increased substantially by h130 bn and repayment was extended over a longer period of time. To implement the PSI in debt restructuring, a cut of 21% of the nominal value of Greek bonds and re-profiling of maturities was decided upon with the tacit agreement of major European banks.

Crucially, the EU authorities finally recognised the perils of recession and allowed Greece to withdraw a total amount of h17 bn from structural funds without applying the fiscal brake of national co-financing. The plan looked



Greek Crisis in Perspective: Causes, Illusions and Failures, Fig. 9 Actual and planned proceeds from privatisations (Note: Proceeds are net of capitalisations in state-owned enterprises. For 2008, 2009 and 2010 figures of proceeds are net of bank shares purchases, thus the negative sign. The recapitalisation of banks after the 2012 PSI was financed by EFSF and is not included here. Source: Government Budget Reports, Ministry of Finance, various years.)

powerful, except for the typical implementation lags. The agreement was voted through by all member state parliaments only in late September 2011 and the release of structural funds was approved by the European parliament in November that year. Participation in the PSI had reached only 70% of institutional holders amid speculation that post-agreement buyers of Greek debt from the heavily discounted secondary market were expecting a huge profit through their offer to cut it!

Thus a new intervention looked inevitable, and in October 2011 a revised debt restructuring plan (the so called PSI +) was authorised, envisaging cuts of 50% of nominal bond value that would eventually reduce Greek debt by h120 bn and would allow it to be stabilised at around 120% of GDP by 2020. In exchange, Greece would undertake further fiscal cuts of around 6% of GDP. Greek bonds held by public institutions (i.e. by the bailout providers) would be fully honoured, though social security funds and domestic public entities were forced to participate in the scheme.

The agreement was hailed by the government and the bailout institutions as the definite solution to the Greek debt conundrum, but the euphoria turned sour a few days later when the government surprised everybody by seeking a referendum for its approval. Many feared that the outcome could in all probability be negative as an expression of current misgivings, and this would be quickly interpreted as opting for exiting the eurozone. In the ensuing furore, the decision was annulled, the Prime Minister resigned and a coalition government was formed in November 2011 to implement the restructuring of debt and negotiate the terms for the new round of EU-IMF loans. The three-party government acted quickly and concluded the PSI agreement in May 2012, but the extra fiscal package was not finalised, as new elections were called. The coalition partners were anxious to refresh their mandate before the political cost of further cuts becomes too inhibiting for them to remain in power. A caretaker government followed, under which privatisations were adjourned so as not to excite opposition from the unions, while pre-electoral inaction made collection authorities slacken the processing of income tax statements, so as not to infuriate voters by presenting them with an increased tax burden.

The situation was further aggravated after the first election round proved inconclusive, with the two mainstream parties seeing their share of the vote collapsing from around 80% of the electorate in 2009 to just one third of the total vote in 2012. Meanwhile, left-wing and far right-wing parties with strong rhetoric against the bailout agreements saw their share of the vote soar; this prompted a new wave of speculation that Greece would be likely to exit the eurozone, and the ensuing capital flight drained significant amounts from the Greek banking system. The second round was polarised by the euro dilemma, thus helping to increase mobilisation of voters and finally resulting in a tripartite coalition that vowed to take all necessary measures to safeguard Greece in the eurozone.

Although the new government exhibited some of the weaknesses typical of party coalitions, it strived to persuade domestic and European opinion that it meant business. As a confirmation signal, it reaffirmed that privatisation of several public companies would go ahead and announced that the state would abolish minimum holding rights in utilities so that potential buyers would have more incentives to increase their offers. The Agricultural Bank was swiftly privatised and other state-owned credit organisations were chosen to follow suit. The fiscal package was finally approved by the Greek parliament in autumn 2012, through a new round of social tensions and partisan breakaways by disillusioned members.

But, in the meantime, two factors had adversely affected the situation of Greece versus the eurozone and the bail-out authorities: on the European front, the market pressures were directed towards the economies of Spain and Italy, and the concomitant threat to the very existence of the euro pushed the Greek problem to the sidelines. Although European authorities responded to the challenge by designing a new defence mechanism in the banking system, and the ECB decided to intervene in the bond markets to stave off speculative attacks against member

states, European public opinion was overwhelmed by 'rescue fatigue' and appeared to be increasingly hostile towards granting any further support for Greece.

On the other hand, Greece itself was already overstressed: unemployment was rocketing above 24% of the labour force, while recession deepened further with the contraction of real GDP reaching 6.7% in 2012 and another 3.9% in 2013, a huge deterioration from milder estimates when the PSI programme was approved. The deterioration was finally halted in 2014, but the anaemic growth of 0.77% hardly convinced anybody that recession was just about to vanish.

As a result of such developments, debt stabilisation was completely jeopardised and a new cycle of futility and desperation emerged leading up to the collapse of the government in the 2015 elections. Since the dynamics of the debt-to-GDP ratio are sensitive to the prospects of growth, it is worth examining alternative debt paths that correspond to higher growth profiles.

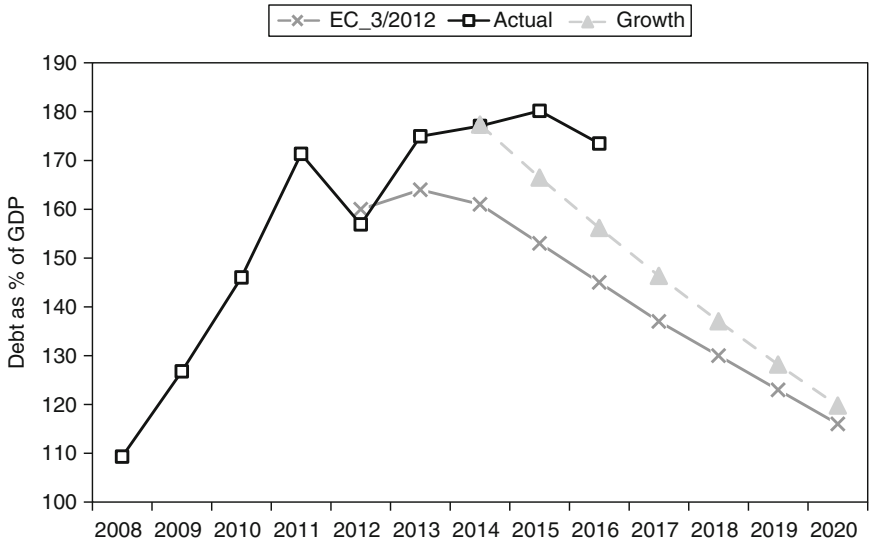
An Alternative Scenario: Walking on a Tight Rope

A benchmark for debt developments is the projection over the period 2012–2020 that has been calculated by European authorities in March (European Commission 2012, Table p. 30) to ensure sustainability of public debt after the PSI. This was based on real growth resuming in 2014 and staying above 2% thereafter, while inflation of the GDP deflator is consistently kept below 2%. A primary budget surplus of over 2% of GDP was foreseen for 2013 and assumed to remain above 4% of GDP for the rest of the period. Privatisations were forecast to be extensively implemented and, as explained above, the target set by the government for 2014 was to collect h3,330 million, or 1.9% of GDP. However, the actual outcome was far from all these optimistic projections: the primary surplus in 2013 and 2014 was below and slightly above 1% of GDP respectively, while privatisation proceeds in 2014 were only h584 million, i.e. roughly one sixth of the amount initially envisaged. Furthermore, the economy continued to suffer from deflation, thus implicitly augmenting the debt to output ratio.

As a consequence, debt continued to rise further and in 2014 reached 177% of GDP, well above the 161% projected by the European authorities, as shown in Fig. 10. To understand the primary role of growth in returning the debt burden to a viable trajectory, an alternative scenario is constructed based on a number of assumptions:

1. The primary surplus is set at 2% of GDP. Previous targets of surpluses at 4% of GDP per annum are wholly unrealistic. Even if they were to be achieved by further intensifying the austerity programme, they would cause deep recession and debt as a ratio to GDP would rise further.
2. Real growth rate is set at 3% per annum; this can be achieved by exploiting the fiscal multiplier effect of easing the austerity programme and boosting investment.
3. Inflation rate at 2%, which is compatible with a relaxation in fiscal consolidation and the return of growth.
4. Privatisation proceeds are set at 1% of GDP.
5. The average cost of borrowing is set at 1% per annum; this can be achieved through an agreement to reduce the cost of debt servicing and/or may also include rescheduling of repayment obligations.

The alternative debt profile is shown in Fig. 10. The higher growth scenario leads to debt to output ratio constantly declining and approaching a level of around 120% at the end of the period. The growth rate assumption is not unrealistic: growth rates are still lower than those that prevailed in the previous decade. Possibly they could be further enhanced by implementing proper market reforms and inviting more private investment. Other assumptions, such as those of medium primary surpluses and systematic privatisations, also become more realistic with higher growth. Sustainability may be further assisted if the bailout funds currently allocated for the recapitalisation of Greek banks are taken out of public debt accounts and become



Greek Crisis in Perspective: Causes, Illusions and Failures, Fig. 10 Alternative paths for public debt as % of GDP. (Note: Actual data are taken from Ameco as in Fig. 1. The figure for 2015 is a prediction based on the official fiscal targets. The EC path is taken from the post-PSI Report EC (2012) and is based on high primary surpluses and low growth rates. The 'Growth path' is constructed by the author on the following assumptions: (i) Real growth rate 3% per annum; inflation rate 2%; primary surplus 2% of GDP; privatisation proceeds at 1% of GDP, and average cost of borrowing 1% per annum.)

liabilities of the new banking authority that is scheduled to operate next year on a eurozone-wide basis. This will mean a further reduction of the debt to output ratio by more than 25 percentage points and will firmly anchor Greece in the eurozone.

Is Exiting from the Euro an Option Worth Considering?

The crisis in Greece has had profound ramifications for the eurozone, both in political as well as in economic terms. In the euro area, Greece is routinely considered not only as devouring European taxpayers' money, but also as the habitual wrongdoer, especially when compared with the other three countries (Ireland, Portugal and Cyprus) that underwent similar – though not as recessionary – adjustment programmes with higher efficacy. In such a politically unyielding and increasingly suspicious framework, a Greek exit from the eurozone has started to attract attention both at home and abroad.

Though the complications and costs that will ensue in the Eurozone banking sector can be enormous, the exit of Greece could prove opportunistically attractive to some European politicians, who get angrier every time a new round of aid is discussed. However, they overlook the fact that a Greek exit would reverberate around other states and lead to an aggravation of the crisis; for how contagion will spread, see Vehrkamp (2011). It may also serve as a convenient argument for consolidating and enforcing a two-tier model of economic governance, as was advocated before the creation of EMU (e.g. Bayoumi and Eichengreen 1992). The idea is recently suggested again by commentators and politicians betting on the 'Grexit scenario' and assuming that other countries may follow suit. Based on an inner core of surplus economies in the north and a weaker periphery in the south, competitiveness in this model will be restored through the so called 'internal devaluation' of labour costs, thus perpetuating the gap that is already widening between the eurozone countries; for a description of divergences within the common currency see Christodoulakis (2009).

For Greece, exit would trigger a prolonged economic catastrophe. As the entire Greek debt will remain denominated in euros, the rapid depreciation of the new national currency will make its servicing unbearable and the next move will be a disorderly default. Isolation from international markets would drive investors even further away, while the financial panic would drain domestic liquidity at a massive scale. The creditor countries of the EU would start demanding repayment of their aid loans, and this would soon deprive Greece of its claim on the EU cohesion funds. Tensions would be likely to produce further conflicts with EU agencies and the pressure to consider complete disengagement from the EU would gain momentum both domestically and abroad.

Stay in the Eurozone and Grow More

The cost would be so immense that the single available option for Greece is to complete the fiscal adjustment and become reintegrated into the eurozone as a normal partner. This requires Greece to undertake concrete actions that produce visible results within a short time frame, so that society becomes more confident to pursue further reforms. Some policy suggestions for this direction are as follows.

First, Greece needs to acquire credibility while its problems are being properly understood abroad. The continuing fiscal shortfall is easily translated as reluctance, causing continual friction with the EU and demands for a new battery of austerity measures. To escape this cycle, Greece must adopt a front-loaded policy as a matter of urgency to achieve key fiscal targets quickly and to change the impression of being a tactical waver. This line was initially adopted by the previous government, but was later abandoned as the coalition was reluctant to inflict costs on various pressure groups and pay the price in the polls. The new government is still indecisive as to where the burden of adjustment is likely to be felt more.

If the government finally decides to implement such a front-loaded policy, it may be in a position to revise some of the pressing – although so far unattainable – schedules set by the creditors and thus ensure higher social approval and tolerance. To insulate itself from the risk of a spending spree in future elections, the best option for Greece is to adopt a constitutional amendment on debt and deficit ceilings, just as Spain did in 2011, alleviating market pressures, at least for the time being.

Second, Greece desperately needs a fast-track policy for exiting the long recession. A substantial amount could be allocated from the so-called 'Juncker investment plan' that is currently promoted by the EU to support major infrastructural projects and private investment in export-oriented companies. This growth boost should then be followed by structural reforms and privatisations that can attract significant private investment as market sentiment is restored. In addition, instilling growth will help to control the debt dynamics and reduce public deficits without ever-rising taxes that thwart private investment and make economic recovery and sustainability even less attainable. Feldstein (2012) leaves no doubt about the mechanics of stabilisation when he warns that

> "[t]o achieve a sustainable path, Greece must start reducing the ratio of its national debt to GDP. *This will be virtually impossible as long as Greece's real GDP is declining*", (my emphasis).

The inevitability of the above thesis can no longer be ignored. Nor can it be circumvented by sermons on the necessity of front-loaded reforms on the assumption that they will automatically restore growth and competitiveness.

Conclusions

Exactly three decades after becoming a fully fledged member of the EU and ten years after joining the eurozone, Greece sought a bailout agreement in 2010 to avoid bankruptcy. A long history of stabilisation programmes proved incapable of achieving a lasting fiscal correction and adequately raising competitiveness, as fundamental weaknesses in the economic and political system continued to play a corrosive role. The oversized public sector and the frequent indulgence in pre-electoral spending sprees in exchange for

political support led to protracted fiscal deficits and the accumulation of a large public debt.

Equally, the chronic deterrence of productive investment by a multitude of regulatory inefficiencies resulted in a thin tradeable sector and large current account deficits. The economy remained vulnerable to political developments which were often dictated by short-term partisan considerations with far-reaching fiscal implications. This explains why, in spite of substantial reforms taking place over the last two decades and achieving high growth rates, EMU participation and moderate debt stabilisation, the situation went once more out of control.

Regarding the current crisis, the article has described how prolonged external and fiscal deficits were allowed to reach uncontrollable levels and, in the aftermath of the credit crunch, led to a further escalation of debt and the subsequent bailout. Five years later, fiscal consolidation is still far from being sustainable in spite of augmenting the bailout loans and implementing a substantial debt reduction on private holders. The economy has contracted by nearly 25% since 2008, social tensions have been multiplying and eventually elections were won by a coalition that pledged to dismantle the austerity programme.

The post-election domestic and European ambivalence over the course that economic policy should follow produced such uncertainty worldwide that the future of Greece in the eurozone was once again put in jeopardy. Some consider such an outcome as a due punishment for past excesses, while others see it as an escape from further unemployment and recession.

Events took a new twist in July 2015. After months of inconclusive negotiations with a multitude of euro area authorities, the Greek government surprised everybody by leaving the talks and announcing a referendum on the latest version of the lenders' proposals. Amid strong controversies on the real nature and meaning of the vote, the result was a sound rejection of the new austerity framework. Frightened by the prospect of Grexit in case talks break down, the government interpreted the outcome as a strong incentive to embark on a new round of negotiations demanding a policy mix with less austerity and some unspecified growth initiatives, capped with a pledge to re-examine the debt issue.

The article finds both points of view to be inadequate, and argues that the only viable way out of the current crisis is to strike an agreement with creditors that involves measures to restore growth, accompanied by realistic targets for privatisations, fiscal consolidation and market reforms.

The lesson of the past five years is that deep recession will otherwise continue to hinder any existing possibility for exiting the crisis. Greece, and other eurozone countries too, are desperate for a 'corridor of confidence', to use Keynes' famous phrase, to put things in order before it is too late for both Greece and Europe as well.

See Also

- ▶ Euro Zone Crisis 2010
- ▶ European Central Bank and Monetary Policy in the Euro Area
- ▶ European Central Bank
- ▶ European Cohesion Policy
- ▶ European Monetary Integration
- ▶ European Monetary Union
- ▶ European Union (EU) Trade Policy
- ▶ European Union Budget
- ▶ Genuine Economic and Monetary Union
- ▶ Germany in the Euro Area Crisis
- ▶ Stability and Growth Pact

Acknowledgments I have benefited from various comments by V. Sarantides and A. Ntantzopoulos and I am thankful to participants in a LSE seminar on an earlier version of this article.

Appendix: A Brief Description of the Conditionality Programmes for Greece

The adjustment programme for Greece was laid out in three phases. The first Memorandum was signed in May 2010 and aimed at reducing the fiscal deficit to 3% in 2013. Specific measures that were actually implemented included universal cuts in public salaries and all pensions, a rise in VAT from 19% to 23% and similarly in other

consumption surcharges, the abolition of collective agreements in favour of firm-level contracts, the lowering of private sector wages by 12% and a reform in the social security system. It also included the liberalisation of red tape practices in the transport, pharmacy and legal sectors, but the outcome was heavily compromised through a series of delays and back-offs. The fiscal deficit for 2010 ended up close to 11% of GDP, substantially lower than the horrendous 15.4% in the year before, but still away from the initial target.

Thus in early 2011 a new round of negotiations resulted in a second round of measures voted for by parliament in June 2011. They included further taxation on past incomes, a lump-sum tax on professionals, further rises in indirect taxes and a new property levy that was imposed two months later. The programme demanded the abolition of outdated public entities, a reduction in the number of civil servants and a further curtailment in their salaries. It also envisaged ambitious privatisations of utilities and public real estate that could trim down public debt by h50 bn within a four-year period. The fiscal deficit for 2011 was estimated to be 9.8% of GDP, revealing a major difficulty in further adjustment in the absence of growth.

The third round of adjustments was voted for in February 2012 as Memorandum II. (For the full text see 'Memorandum of Understanding on Specific Economic Policy Conditionality', 9 February 2012, available at http://www.hellenicparliament.gr).

This time it was approved by the two major parties, but only after a line-up was imposed to avoid desertions and rising internal protest. Measures included a reduction of minimum wages in the private sector by 22%, an additional cut by 10% to new entrants as a means to beat youth unemployment, 15% cuts in various pensions, the abolition of several tax credits and explicit targets for cutting employment and entities in the wider public sector. Policies started to be implemented in the final quarter of 2012. After the referendum in July 2015, a new round of measures was agreed to be voted by Greek Parliament. The process is expected to be completed in the autumn 2015 and a third Memorandum will apply for the next three years.

Bibliography

Bayoumi, T., and B. Eichengreen. 1992. Shocking aspects of European monetary unification. *CEPR Discussion Paper* No. 643, May.

Blanchard, O. 2006. *Current account deficits in rich countries*. IMF Mundell-Fleming Lecture.

Blanchard, O., and F. Giavazzi. 2002. Current account deficits in the euro area: The end of the Feldstein–Horioka puzzle? *Brookings Papers on Economic Activity* 33: 147–210.

Christodoulakis, N. 1994. Fiscal developments in Greece 1980–93: A critical review. *European economy: Towards greater fiscal discipline*, No. 3, 97–134. Brussels.

Christodoulakis, N. 2009. Ten years of EMU: Convergence, divergence and new priorities. *National Institute Economic Review* 208: 86–100.

Christodoulakis, N. 2010. Crisis, threats and ways out for the Greek economy. *Cyprus Economic Policy Review* 4(1): 89–96.

Christodoulakis, N. 2012. Market reforms in Greece 1990–2008: External constraints and domestic limitations. In *From stagnation to forced adjustment: Reforms in Greece, 1974–2010*, ed. S. Kalyvas et al., 91–116. New York: Columbia University Press.

Christodoulakis, N., and V. Sarantides. 2011. External asymmetries in the euro area and the role of foreign direct investment. *Bank of Greece Discussion Paper*.

De Grauwe, P. 2010. The Greek crisis and the future of the eurozone. *Intereconomics* 2: 89–93.

European Commission. 2009. *Quarterly Report on the Euro Area*. 8(1). Brussels.

European Commission. 2011. The economic adjustment programme for Greece: Fifth review. *European Economy*, Occasional Papers 87, October.

European Commission. 2012. The second adjustment programme for Greece: Fifth review, *European Economy*, Occasional Papers 94, March.

Feldstein, M. 2012. The failure of the euro: The little currency that couldn't. *Foreign Affairs* 91(1): 105–116.

Flood, R., and C. Kramer. 1996. Economic models of speculative attack and the drachma crisis of May 1994. *Open Economies Review* 7: 591–600.

IMF. 2011. Greece: Third review under the stand-by arrangement. *Country Report* No. 11/68, March, International Monetary Fund.

Krugman, P. 2011. Origins of the euro crisis. http://krugman.blogs.nytimes.com/2011/09/23/origins-of-the-euro-crisis/. 23 September. Accessed 7 July 2015.

Memorandum II. 2012. *Memorandum of understanding on specific economic policy conditionality*, 9 February. Available at http://www.hellenicparliament.gr

Shelburne, R.C. 2008. Current account deficits in European emerging markets. *UN Discussion Paper* No. 2008.2.

Skouras, S., and N. Christodoulakis. 2014. Electoral misgovernance cycles: Tax evasion and wildfires in Greece. *Public Choice*, 159(3–4): 533–559, June. Available at http://link.springer.com/journal/11127/159/3/page/1

Vehrkamp, R. 2011. Who's next? The eurozone in an insolvency trap. *Bertelsmann Stiftung*, No. 2.

Greek Crisis in Perspective: Origins, Effects and Ways-Out

Nicos Christodoulakis
Department of International and European Economic Studies (DIEES), AUEB, Athens University of Economics and Business, Athens, Greece

Abstract

In 2011 the Euro faced its toughest challenge since its introduction as several of the participating Member States faced unprecedented financial problems. Greece was the most severe case requiring intervention from the EU and IMF to stabilize its economy and repay debt obligations. This article explains the debt process in Greece from the 1980s to date, and describes its main causes and episodes. It also assesses the IMF-EU Memorandum and argues that the collapse of growth inhibits the prospects of debt stabilization. An alternative scenario is discussed showing that stabilization can become more effective and realistic if recession is tackled first and reforms follow on a steadier path.

Keywords

Debt; Fiscal deficits; External balances; Crisis; Eurozone; Greece

JEL Classifications

H60; H61

Introduction

In the aftermath of the global financial crisis of 2008, a number of Eurozone countries were engulfed in a spiral of rising public deficits and explosive borrowing costs that eventually drove them out of markets and into bail-out agreements jointly undertaken by the International Monetary Fund (IMF), the European Union (EU) and the European Central Bank (ECB). Greece was by far the most perilous case with a double-digit fiscal deficit, an accelerating public debt which in GDP terms was twice as much the Eurozone average and an external deficit near 5,000 US Dollars per capita in 2008, one of the largest worldwide. No wonder that Greece was the first to seek the bail-out assistance and the last expected to exit its ever-changing conditionality terms.

Two and a half years after the bail-out Memorandum was signed, the situation remains highly uncertain. The economy faces an unprecedented recession, unemployment is rocketing, social unrest undermines the implementation of reforms and the fiscal front is not yet under control, despite extensive cuts in wages, salaries and pensions. In the summer of 2011 uncertainties multiplied at such a rate that the possibility of Greece exiting the Eurozone was widely discussed either as a punishment mechanism from abroad for not accepting the pains of adjustment or as a quick fix from within to avoid them for good. In two subsequent EU summits, held respectively in July and October 2011, the Memorandum agreement was substantially broadened to include a radical debt reduction, a second round of bail-out loans by IMF and the EU and a generous release of European structural funds to assist the real economy. The agreement was conditional on being approved by the national Parliaments of the lender states as well as by the European Parliament. Finally, the conditionalities of the Memorandum were approved by the Greek Parliament in February 2012 and the debt-cutting process was concluded in May. However, most of the envisaged measures were delayed for the third quarter of the year, as two round of elections took place to provide new legitimacy for carrying on the program and implementing reforms. The prolonged electoral uncertainty meant that most of the adjustment measures were weakened or postponed, leading to new tensions over Greece's determination. A coalition Government was finally formed in June by parties vowing to apply all policies deemed necessary for the country to remain in the Eurozone, though at the same time seeking some relaxation of the time frame from European authorities.

It is tempting to note that the economic capacity of the country to adjust and the social endurance are diminished exactly when the European environment is becoming more helpful for stressed countries, This makes the Greek problem an unusually interesting case for analysis, not only for understanding its origins and causes but also for devising a realistic strategy to solve it.

The purpose of the present article is twofold: First to provide a historical account of debt accumulation, identify the main difficulties of fiscal stabilization and explain the factors that led to the present crisis and the failure to prepare for it. Second, to assess the main reasons for missing the targets set by the Memorandum agreement and the need for encompassing a growth strategy in order to make reforms acceptable and more effective to achieve debt sustainability in the longer run.

Section "The Period of Debt Escalation: 1980–1993" describes the main episodes of debt escalation in the 1980s, section "Debt Stabilization and EMU Membership" the stabilization effort on the way to EMU and section "Unprepared for the 2008 Crisis" the toxic combination of fiscal irresponsibility, external deficits and political indecision during the more recent period that led to the present crisis. Section "Two Important Policy Facts" describes some recurrent facts on fiscal policies that repeatedly hinder stabilization and growth. Section "An Ex Post Assessment of the Memorandum" attempts an ex post assessment of the policies conditioned by the Memorandum agreement to correct the economy while section "The New Memorandum Conditionalities and Ways-Out of the Crisis" argues why exiting the Eurozone should not be an option for Greece. An alternative scenario based on higher growth is shown to be more credible in achieving fiscal consolidation and stabilizing the debt over the medium term. Section "Conclusions" concludes with the need to fight current recession as the only way for Greece to regain social coherence and debt sustainability in the new landscape of the Eurozone.

The Period of Debt Escalation: 1980–1993

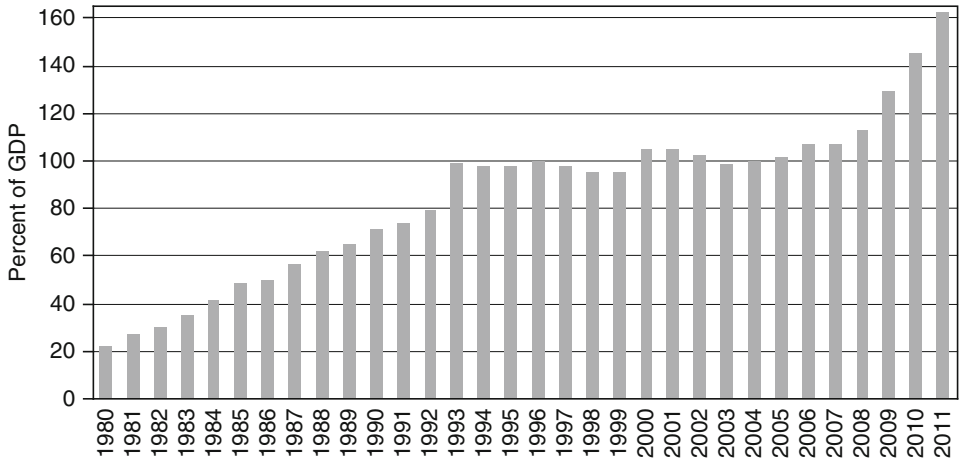
In 1980, Greece became a full-fledged member of the European Union and this marked a wholly new period for the economic and political developments in the country. Greece was one of the first non-founding countries to start accession talks with the Common Market as early as 1961, but the process was abruptly suspended with the advent of the military dictatorship that lasted until 1974. Membership of the European Union was rightly viewed as an anchor of political and institutional stability for the newly restored democracy, but nonetheless it also fed and multiplied uncertainties over the economy.

After a long period of growth, Greece entered a period of recession in late 1970s, not only as a consequence of worldwide stagflation, but also because – on its way to integration with the common market – it had to dismantle its preferential system of subsidies, tariffs and state procurement by which several companies were kept profitable without being competitive. Soon after accession, many of these companies went out of business and unemployment rose for the first time in many decades.

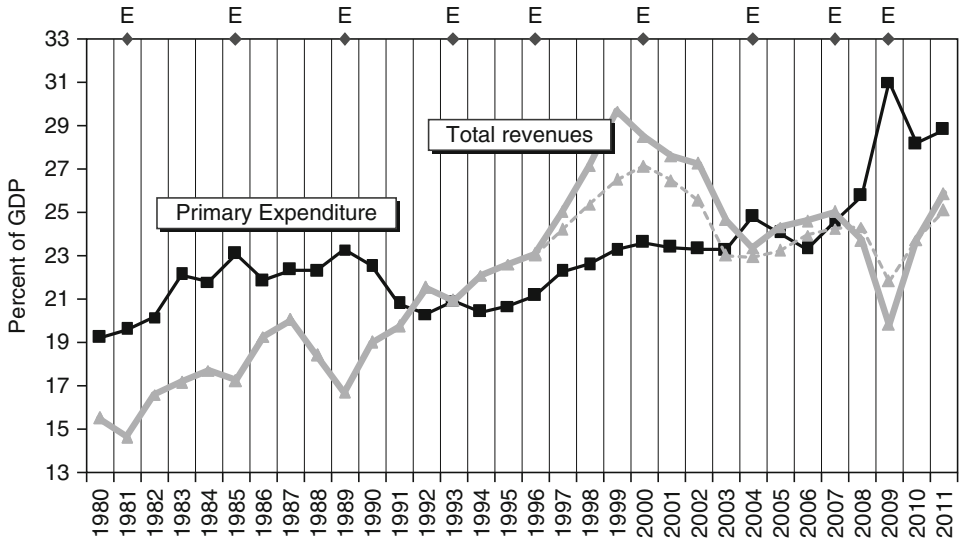
The Government opted for a massive fiscal expansion that included demand–push policies to boost activity and the public underwriting of several ailing companies to maintain employment. The effect was quite predictable: private debts turned into a chronic hemorrhage of budget deficits without any supply-side improvements. Similarly, the expansion of demand simply led to more imports and higher prices. Activity got stuck and Greece ended up in a typical stagflation, perhaps the quickest assimilation to European practices of the time.

As a result, accession to the Promised Land strangely coincided with the unleashing of a nightmare thought to be in dormant thus far: *public debt*.

Looking at Fig. 1, there are three distinguishable phases for the dynamics of debt: The first covers the period 1980–1993 during which public debt rose from slightly above 20% of GDP toward 100% in 1993. The second phase spans the period



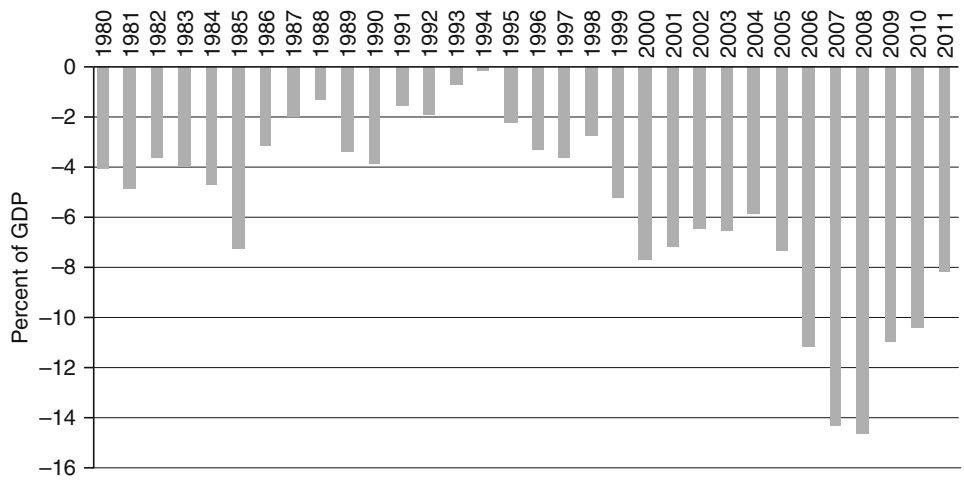
Greek Crisis in Perspective: Origins, Effects and Ways-Out, Fig. 1 Greek public Debt as %GDP for the period 1980–2011 (Source: Debt of General Government, ESA95 definition, Ameco Eurostat 2011. GDP at market prices, IMF WEO Database 2010)



Greek Crisis in Perspective: Origins, Effects and Ways-Out, Fig. 2 Primary public expenditure and total revenues (incl. privatisation proceeds) as %GDP in Greece, 1980–2011. Election years denoted by (E). *Dotted line* denotes public receipts net of privatizations (Source: Budget Reports. GDP at market prices, AMECO Database 2012)

1994–2005 in which public debt ends up again at around 100% of GDP after two mild reductions in between. The third phase covers the period 2006–2011 when public debt surpasses the 100% threshold, accelerates after 2008 and ends up exceeding 160% of GDP in 2011.

The above periodicity broadly coincides with substantial shifts in the context of economic policies, as suggested by developments in the fiscal patterns shown in Fig. 2 and in the Current Account depicted in Fig. 3 and briefly discussed below.



Greek Crisis in Perspective: Origins, Effects and Ways-Out, Fig. 3 Current Account in Greece as % of GDP, 1980–2011 (Source: IMF WEO Database 2011)

Regarding fiscal developments, the main characteristic of the first period was the substantial expansion of public spending and the concomitant rise in budget deficits and government debt. Revenues increased as a proportion of GDP, but were outpaced by the steadily growing expenditure. Both fiscal components appear to be volatile in the election years 1981, 1985 and 1989, suggesting the presence of a strong political cycle in public finances, as will be discussed later in more detail.

To maintain competitiveness, authorities had adopted, since the mid 1970s, a real exchange-rate target with a crawling peg. After the Government adopted an automatic wage indexation scheme in 1982, the only effect of the exchange rate policy was to fuel price increases and aggravate trade deficits. To break the vicious cycle of depreciation and inflation, a discrete devaluation combined with a temporary wage freeze was implemented in 1983, but it was superseded by a new phase of expansion as elections were approaching leaving public debt at even higher levels.

The external deficit approached 8% of GDP in 1985, an alarming threshold as several Latin American economies with similar imbalances were serially collapsing at that time. A coherent stabilization program was called for in October 1985 enforcing a discrete devaluation by 15%, a tough incomes policy and extensive cuts in public spending. The program achieved a rise in revenues by beating several taxevasion practices and replacing previous indirect taxes with the more effective VAT system adopted by the European Union. Public debt was stabilized, but only until the program was finally abandoned in 1988, after being fiercely opposed from within the Government and the ruling party.

The First Fiscal Crisis

Two general elections in 1989 failed to secure majority, thus leading to the formation of coalition Governments, an event that was hailed as a confirmation of political maturing and an opportunity to overcome partisan differences on major issues. But self-indulging admiration was short-lived, as stabilization policies are notoriously difficult to implement through party coalitions because each party tries to avoid the cost falling on its own constituency. Greece was no exception to the rule and the economy suffered a major setback in 1989, far more serious than previous fiscal failures.

Two episodes are characteristic of how a rhetoric designed to please everybody in combination with naïve policies can lead to disaster: Despite looming deficits, in 1989 the coalition Government decided to abolish prison terms for major tax

arrears hoping to induce offenders to repent and reconsider their strategy. Expectedly, the move was interpreted the other way around as a signal of relaxed monitoring in the future, thus encouraging further evasion.

Another bizarre policy was to cut import duties for car purchases by repatriates returning to Greece after the collapse of the Soviet Union. The measure was viewed as a gesture to facilitate mobility back in the motherland, but it was quickly turned into a black-market scheme. At a small bribe, immigrants were purchasing luxury cars only to immediately resell them to rich clients who could thereby avoid the duty tax. The Budget was deprived of badly needed revenues and evaders had yet another reason for celebration.

As a result, revenues collapsed and the country suffered a major fiscal crisis, until a majority Government was elected in 1990 and enacted a new stabilization program. Despite substantial cuts in spending and a rise in revenues, public debt as a ratio to GDP continued to rise because of the higher cost of borrowing worldwide and a stagnant output. The sharp rise in 1993 in particular, is due to the inclusion of extensive debts initially contracted by public companies under state guarantees but finally underwritten by the Budget. Except for the electoral years 1989–90, fiscal consolidation significantly improved the Current Account and such a rarity as a balanced external position was reached in 1994.

Debt Stabilization and EMU Membership

Although Greece was a signatory of the Maastricht Treaty in 1991, it was far from obvious whether, how and when the country could comply with the nominal convergence criteria required to join the Economic and Monetary Union. Public deficits and inflation were galloping at two-digit levels and there was great uncertainty about the viability of the exchange rate system; for a detailed analysis of the period see Christodoulakis (1994).

In May 1994, capital controls were lifted in compliance with European guidelines and this prompted fierce speculation in the forex market. Interest rates reached particularly high levels and the Central Bank of Greece exhausted most of its reserves to stave off the attack; for an account of the successful defense see Flood and Kramer (1996). This episode proved to be a turning point for the determination of Greece to pursue accession to EMU in order to be shielded by the common currency and avoid similar attacks in the future. Soon afterwards the "Convergence Program" was adopted that set time limits to satisfy the Maastricht criteria and included a battery of reforms in the banking and the public sectors.

International markets were not impressed and continued to be unconvinced about exchange rate viability. With the advent of the Asian crisis in 1997 spreads rose again dramatically and – after months of credit shortages – Greece finally decided to devalue by 12.5% in March 1998 and subsequently enter the Exchange Rate Mechanism wherein it had to stay for two years. The country was not yet ready to join the first round of Eurozone countries in 1998,and Greece was granted a transition period to comply with the convergence criteria by the end of 1999.

After depreciation, credibility was further enhanced by structural reforms and reduced state borrowing so that when the Russian crisis erupted in August 1998, the currency came under very little pressure. Public expenditure was kept below the peaks it had reached in the previous decade and was increasingly outpaced by the rising revenues and various one-off receipts. Tax collection was enhanced by the introduction of a scheme of minimum turnover on SMEs, the elimination of a vast number of tax allowances, the imposition of a new levy on large property and a rereorganization of the auditing system. Proceeds were further augmented by privatization of public companies and, as result, public debt fell to 93% of GDP in 1999. Although still higher than the 60% threshold required by the European Treaty, Greece benefited from the convenient interpretation that it suffices "*to lean toward that level*", as previously used by other countries – such as Italy and Belgium – in their own way to enter EMU.

The Implementation of Market Reforms

In the 1980s, structural reforms were hardly on the agenda of Greek economic policy. In fact for most of the period the term was a misnomer used to describe further state intervention in economic activity, rather than market-oriented policies as practised in other European countries. Market reforms were introduced for the first time in 1986 aiming at the modernization of the outmoded banking and financial system in compliance with European directives. A major reform in social security took place for the first time in 1992, curbing early retirement and excessively generous terms on the pension/income ratios.

Throughout the 1990s, various reform programs were aimed at the restructuring of public companies whose chronic deficits had contributed to the fiscal crisis in 1989. Privatization was attempted through direct sales of state-owned utilities as the quick way to reduce deficits. Despite some initial success, the program was fiercely opposed by the trade unions of public companies and eventually led to the demise of the Government. Privatizations were conveniently brandished as sell-outs, and it took a few more years for the concept to reappear on the political agenda.

A new wave of reforms was launched after 1996 in the course of the "Convergence Program". State banks were privatised or merged, dozens of outmoded organizations were closed down, and a series of IPOs – taking advantage of the stock market bonanza – provided capital and restructuring finance to several public utilities. Other structural changes included the lifting of closed-shop practices in shipping, the entry of more players into the mobile telephony market and a series of efforts to make the economic environment more conducive to entrepreneurship and employment.

Post-EMU Fatigue

After 2000, Greece emulated some other euro area members in exhibiting a '*post-EMU fatigue*' and the reform process gradually slowed down. As shown in Fig. 9, proceeds from privatization peaked in 1999, but subsequently remained low as a result of the contraction in capital markets after the dot.com bubble and the global recession in 2003; for an extensive discussion of reforms in Greece over the period 1990–2008 see Christodoulakis (2012).

An attempt in 2001 to deeply reform the pension system led to serious social confrontations and was finally abandoned. Though replaced by a watered-down version one year later, the failure left a mark of reform timidity for many years. Two other mild reforms followed in 2006 and 2010, but the social security system is still characterised by inequalities, inefficiencies and structural deficits that exert a substantial burden on the General Government finances.

The fatigue spread more widely after the Olympic Games in 2004. With the exemption of the sale of Greek Telecom to the German state company and the privatization of the national air carrier after a decade of failed attempts, most other reforms were consisting of small IPOs with no structural spillovers to the rest of the economy.

Why Debt Reduction Was Insufficient

Despite having achieved substantial primary surpluses throughout 1994–2002 – and around 1999 in particular – public debt over the same period fell only slightly. There are three reasons to explain this outcome. First, during this period the Government had to issue bonds to accumulate a sufficient stock of assets for the Bank of Greece as a prerequisite for its inclusion in the Euro-system, and this capital injection led to a substantial increase in public debt without affecting the deficit.

Second, after a military stand-off in the Aegean in 1996, Greece increased defence procurement to well above 4% of GDP per year. In line with Eurostat rules, the burden was fully recorded in the debt statistics at the time of ordering but only gradually in the current expenditure following the pattern of actual delivery of equipment. This practice created a considerable lag in the debt-deficit adjustment and, in 2004, the Government enforced a massive revision of the deficit figures by retroactively augmenting public spending on the date of ordering, prompting a major dispute over the quality and integrity of the statistics of public finances in Greece. Though a decision by Eurostat in 2006 made the delivery-based rule obligatory for all countries, Greece did not

withdraw the self-inflicted revision. As a consequence, deficits were statistically augmented for 2000–2004 and scaled-back for 2005–2006 relative to what they should have been otherwise, in an awkward demonstration of political interference.

The third reason was the strong appreciation of the Yen/Euro exchange rate by more than 50% between 1999 and 2001. This significantly augmented Greek public debt as a proportion of output due to the fact that substantial loans were contracted in the Japanese currency during the 1994 crisis. To alleviate this exogenous deterioration, Greece entered a large currency swap in 2001 by which the debt to GDP ratio was reduced by 1.4% in exchange for a rise in deficits by 0.15% of GDP in subsequent years, so that the overall fiscal position remained unchanged in present value terms. Although the transaction had no bearing on the statistics for 1999 on which EMU entry was assessed, Greece suffered extensively from criticisms that mistook the swap as a ploy to circumvent a proper evaluation. Values shown in Fig. 1 are net of swap effects, and this partly explains the peak in 2001.

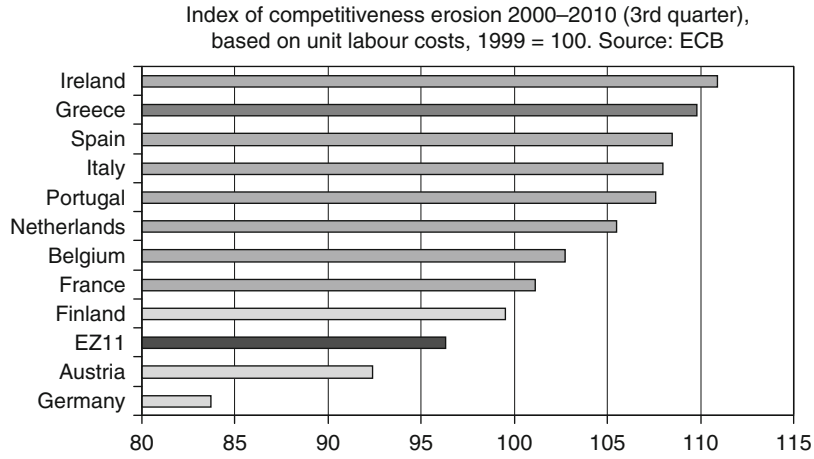
The Current Account

After the Eurozone became operational, hardly any attention was paid to Current Account imbalances, regarding Greece or any other deficit country. Even after they reached huge proportions, external disparities in the euro area continued to remain surprisingly unnoticed from a policy point of view. It was only in the aftermath of the 2008 crisis that policy bodies in the European Union started emphasising the adverse effects that external imbalances may have on the sustainability of the common currency (see for example EC 2009).

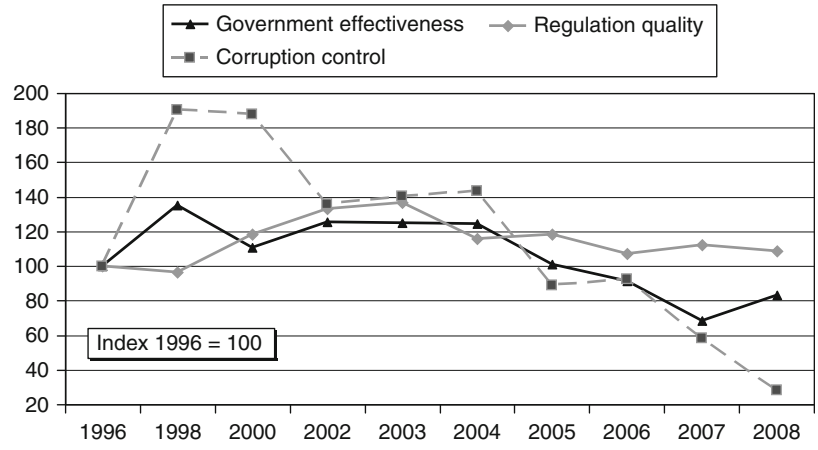
The reason for this complacency was not merely that devaluations were ruled out by the common currency. A widespread – and unwisely comfortable – view held that external imbalances were mostly demand-driven effects and, as such, they would sooner or later dissipate as a result of ongoing fiscal adjustment in member-states. When, for example, Blanchard and Giavazzi (2002) asked whether countries such as Portugal or Greece should worry about and take measures to reduce their Current Account deficits they "... *conclude (d), to a first order, that they should not*". A few years later this proved to be just another misguided assessment; Blanchard (2006) – overturning his previous optimism – remarked that Current Account deficits were steadily increasing within the euro area and urged immediate action otherwise "...*implications can be bad*". And indeed they were.

Although improved for a while after the country joined the common currency, the subsequent vast deterioration in the Greek Current Account played a crucial role in inviting the global crisis home. The reason behind the initial containment was that factor income flows from abroad increased as a result of extensive Greek Foreign Direct Investment in neighbouring countries while labour immigration kept domestic wage increases at bay. The deficit started to deteriorate after 2004 as domestic demand peaked in the post Olympics euphoria, inflation differentials with other Eurozone countries widened and the Euro appreciated further. Unit labour costs increased and as shown in Fig. 4 the relevant index rose by 10% in the period 1999–2010. As an *ex post* wisdom, it is worth noticing from the same figure that a similar erosion of competitiveness took place in *all* other Eurozone countries that are currently in bailout agreements (Ireland by 12% and Portugal 8%) or considered to be at the risk of seeking one (Spain by 9% and Italy by 8%).

Compared to Germany, Greek unit labour costs increased by 27% causing significant bilateral imbalances. However, this erosion was gradual and cannot have been the single reason for the rapid deterioration experienced after 2006. Other factors affecting the investment environment, such as the quality of the regulatory framework, elimination of corruption practices and overall Government effectiveness might as well have been crucial in shaping productivity and competitiveness. Using the Worldwide Governance Indicators published by the World Bank as proxies for how the above factors evolved during the period from 1996 to 2008, Fig. 5 shows that, despite some improvement in the first years of EMU, there was a noticeable decline thereafter.



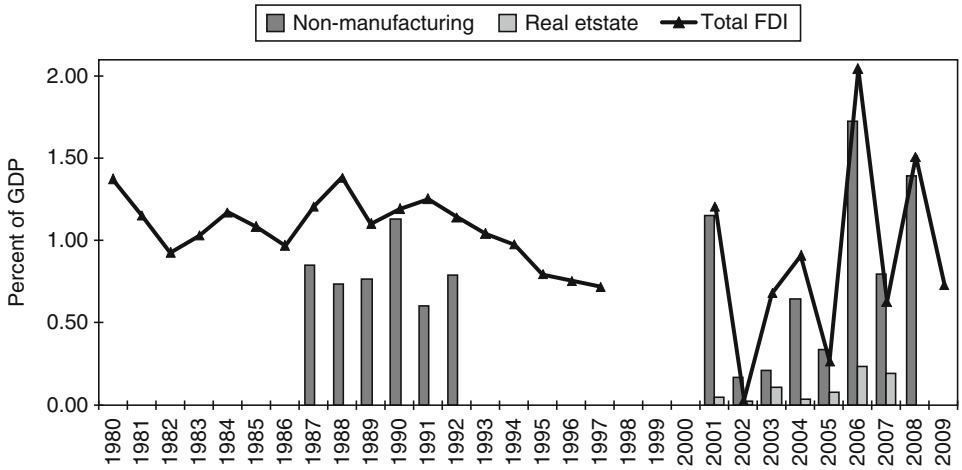
Greek Crisis in Perspective: Origins, Effects and Ways-Out, Fig. 4 Development of unit labour costs in the Eurozone 2000–10 (Source: ECB, Competitiveness indicators, 2011). For Portugal the ULC index was missing for 2010 and replaced by the CPI index adjusted for differences from ULC by using the estimates for 2011. In the more recent editions, the effect of Greek ULC on competitiveness is even less pronounced, due to the wage-cuts implemented in the last quarter of 2010 and through 2011



Greek Crisis in Perspective: Origins, Effects and Ways-Out, Fig. 5 Quality indicators affecting the economic climate. Notes: Indicators are measured in various units with higher values corresponding to better outcomes; to ease comparison all are here indexed at 100 in 1996 (Source: World Bank, WGI various editions)

These developments were pivotal to the poor performance of Greece in attracting foreign direct investment in spite of the substantial fall in interest rates and the facilitation of capital flows within the Eurozone. As depicted in Fig. 6, FDI expressed as percent of GDP hardly improves during the last decade relative to the 1980s. The composition has also changed, as most of the FDI inflows were directed to non-manufacturing sectors and, pointedly, with an increasing allocation to real estate that further aggravates the strain in the Current Account.

It is a well established fact that when new investments are directed mainly to the tradeable sectors this leads to substantial productivity improvements and favours net exports. In contrast, investments going mostly into the real-estate sector boost aggregate demand, raise prices, cause the real exchange rate to appreciate and hinder competitiveness. These developments manifest a major failure of Greece – and for that matter of other Eurozone countries – to exploit the post-EMU capital flows in order to upgrade and expand production; for details see a study by Christodoulakis



Greek Crisis in Perspective: Origins, Effects and Ways-Out, Fig. 6 FDI inflows to Greece expressed as percent of GDP. Note: Missing observations are due to non-availability and do not necessarily imply poor flows (Source: OECD, FDI statistics)

and Sarantides (2011) who use the differentiation in composition and the asymmetry in the volumes of FDI to explain the diverging patterns of external balances in the Eurozone countries.

Unprepared for the 2008 Crisis

The fiscal decline started with the disappearance of primary surpluses after 2003 and culminated with rocketing public expenditure and the collapse of revenues in 2009, as shown in Fig. 2. Revenues declined as a result of a major cut in corporate tax rate from 35% to 25% in 2005 and extensive inattention to the collection of revenues.

Such decisions were making increasingly evident that stabilizing the economy was not a policy priority of the Government, and further actions soon confirmed the assumption: concerned over the rising deficits in 2007, it sought a fresh mandate to redress public finances but – despite securing a clear victory – no such action was taken after the election whatsoever. Only a few months before the global crisis actually erupted, the Government claimed that the Greek economy was "*sufficiently fortressed*" and would stay immune to the reverberations of international shocks. Even after September 2008, the Government was for a long time ambivalent as to whether implement a harsh program to stem fiscal deterioration or to expand public spending to fight off the prospect of recession. A final compromise at the end of the year included a consumption stimulus combined with a bank rescue plan of Euro 5 bn and a pledge to raise extra revenues. The first two were quickly implemented, whilst the latter was forgotten.

Weakened by internal divisions, the Government continued to be indecisive on what exactly to do and, after a defeat in the European elections in June 2009, it opted for yet another general election in October 2009 asking for a fresh mandate to address the mounting economic problems. In practice, the election period turned to be an opportunity for further largesse rather than of preparation on how to contain it. The fiscal consequences were stunning: total public expenditure was pumped up by more than 5 percentage points exceeding 31% of GDP at the end of 2009. (In levels, it exceeded Euro 62 bn, i.e., twice the size in 2003). The rise was entirely due to consumption as public investment remained the same at 4.1% of GDP; details on how public spending was ballooned are given in Christodoulakis (2010).

Total receipts in 2009 collapsed by another 4% of GDP as a result of widespread neglect in collection and the fact that privatization proceeds turned negative since the Government had to finance the emergency capitalization of Greek

banks. The deficit of General Government spiraled and its figure was serially revised from an estimated 6.7% of GDP before the elections to 12.4% in October 2009, and finally widening to 15.4% of GDP by the end of the year. It was only then that European authorities stopped their onlooking attitude and issued a number of warnings against the spending.

Post-Election Inaction

In spite of the gathering storm in the Autumn 2009, the newly elected Government was far from being determined to achieve immediate fiscal consolidation, constrained as it was by its pre-electoral rhetoric that "*money exists*" and its ideological aversion to controlling trade union demands in public enterprises. Trapped in such unrealistic mentalities, the December Budget for 2010 surprised everybody by including an *expansion* of public expenditure and completely *excluding* privatizations, rather than the other way around. Seeing that no appropriate action had been taken to deal with the situation, rating agencies downgraded the economy, this sparked massive credit default swaps in international markets and the crisis loomed.

The problem Greece faced at that time was an acute shortage of financing for the deficit, not yet one of debt sustainability as it later turned out to be. In this regard, a significant opportunity to diffuse the crisis was missed by the Government and European authorities alike. In order to reduce the risk of spillovers to other markets after the credit crunch in 2008, the ECB had invited private banks of Euro member states to obtain low-cost liquidity by using sovereign bonds from their asset portfolio as collateral securitization; see De Grauwe (2010) for a positive assessment of this policy. As a result of this credit facilitation, yields on Treasury bills remained exceptionally low. But instead of borrowing cheaply in the short term as a means of gaining time to redress the fiscal situation, the Government kept on issuing long maturities despite the escalation of costs. This had dramatic consequences on the perception of the crisis by international markets. Commenting on the cost of confusion, Feldstein (2012) aptly notes that:

> "What started as a concern about a Greek *liquidity problem* – in other words, about the ability of Greece to have the cash to meet its next interest payments – became *a solvency problem*, a fear that Greece would never be able to repay its existing and accumulating debt", (my emphases).

Adding injury to misjudgment, the situation was further undermined when the ECB threatened to refuse collateral status for downgraded Greek bonds, hence fuelling fears that domestic liquidity would shrink and precipitating a capital flight from Greek banks. Three months later the rating requirement was dropped for all Eurozone countries, but the damage was no longer reversible. In early 2010, borrowing costs started to increase for both short and long term maturities, Greece had become a front page story worldwide and the count-down began. Despite the belated ECB generosity, the Government was financially exhausted and in April 2010 sought a bailout.

The Role of External Deficits

The global financial crisis in 2008 revealed that countries with sizeable Current Account deficits are vulnerable to international market pressures because they risk having a "sudden stoppage" of liquidity. Recent studies show that highly indebted EMU countries with large external deficits are found to experience the highest sovereign bond yield spreads. Along this line, Krugman (2011) recently suggested that the crisis in the southern Eurozone countries had rather little to do with fiscal imbalances and rather more to do with the sudden shortage of capital inflows required to finance their huge external deficits.

This explains why immediately after the crisis sovereign spreads peaked mainly in economies with large external imbalances, such as Ireland, Spain, Portugal and the Baltic countries, which were under little or no pressure from fiscal deficits; for a discussion of the effects of credit crunch in emerging markets with large Current Account deficits see Shelburne (2008). In contrast, countries with substantially higher debt burdens but without external imbalances, such as Belgium and Italy, experienced only a small increase in their borrowing costs at that time.

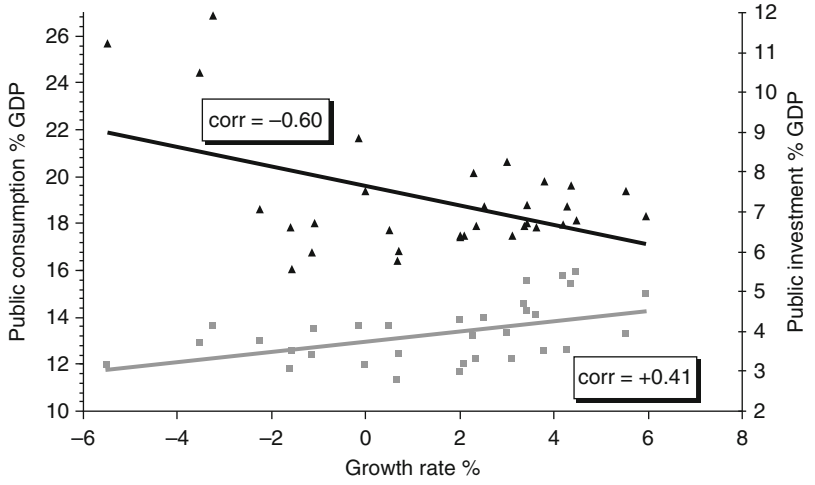
Greece happened to have a dismal record on both deficits and its exposure to the international credit stoppage was soon transplanted into a debt crisis. The Current Account went in free-fall after 2006 when three factors intensified: domestic credit expansion accelerated and disposable incomes were enhanced by the tax cuts, while capital inflows from the Greek shipping sector peaked as a result of the global glut and the huge rise in Chinese freight. The external deficit exceeded 14% of GDP in 2007 and 2008 and still no warning was voiced by any authority, domestic or European. In fact quite the opposite happened: Responding to pleas of car dealers, the Greek Government decided to reduce surcharges on imported vehicles in an attempt to revive the market, while other fellow Governments – at least those from car-making countries – failed to notice the pro-cyclical character of the measure. Replicating history back in 1989, the unfortunate act to facilitate car purchases in order to favour particular groups caused again a significant deterioration of both the external and the public deficit. Additionally, nobody missed the signalling about the true priorities of the Government and the pre-electoral spree followed as described above.

Two Important Policy Facts

Two stylized facts emerge from the historical account of fiscal developments in Greece. One is the fact that in periods of recession counter-cyclical activism usually takes the form of increased consumption, not public investment and this has detrimental effects on public and external deficits without contributing to higher growth. Another recurring characteristic is the propensity of Governments to increase public spending and to tolerate lower revenues in election years.

Cyclicality of Public Spending

As an indication of how the two main components of Government spending behave over the economic cycle, public consumption and public investment expressed as proportions of GDP are correlated with the growth rate; see Fig. 7. Public consumption is found to have a strong negative correlation with growth rates, suggesting a counter-cyclical pattern. This finding implies that periods of economic downturn are likely to be associated with higher public consumption due to increased benefits and programs to contain unemployment. In a situation of fixed public



Greek Crisis in Perspective: Origins, Effects and Ways-Out, Fig. 7 Growth rate correlations with public consumption (Lhs) and public investment (Rhs) expressed as percent of GDP (Source: Government budget, various editions)

employment and nominal wage resistance, public consumption is expected to rise further relative to GDP.

On the other hand public investment shows a strong positive correlation with growth. This implies that, in a downturn, public investment is likely to fall, thus hindering the resumption of growth and causing more recession in the economy.

A clear manifestation of such behaviour over the cycle took place in recent years. With recession deepening year after year, the Government rather than curtailing the public sector found it more expedient to cut public investment in order to control the deficit. As a result, recession was made worse.

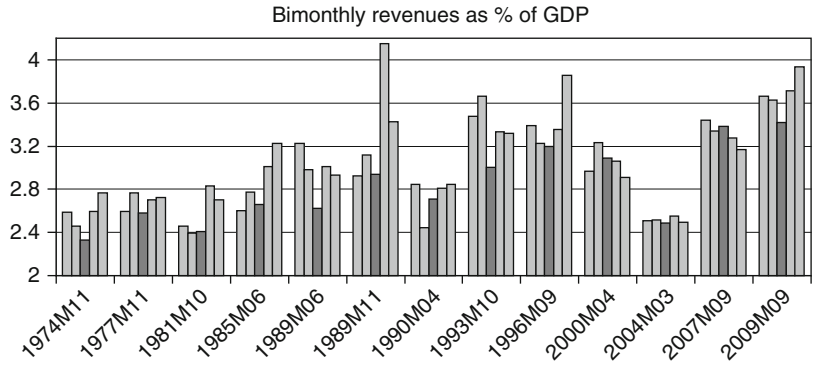
Electoral Cycles

The Greek economy was often subject to the electoral cycle, as incumbent Governments tried to appeal to voters by a variety of opportunistic policies, thus inflicting non-trivial fiscal losses. Practices included extra appointments of party affiliates, grants to favorable groups and allocation of petty projects to local constituencies, all of which affect current or next period expenditure.

It can readily be seen from Fig. 2 that spending rises during the election years in the 1980s and, as deficits widened, the economy had to enter a period of stabilization that was usually terminated before the next election. During the debt escalation in 1980–93 there were four stabilization programs and ten Finance Ministers – usually one to pursue the program and then a successor to denounce it and prepare for the next period of spending rise. Though the electoral cycle subsided in the period before and after EMU membership, it returned full-steam in the elections of 2009.

Apart from direct actions on the expenditure side, the empirical evidence suggests that slacker tax auditing around elections causes further fiscal deterioration. An extensive investigation by Skouras and Christodoulakis (2011) found that flaws in tax collection arise either as a result of deliberate relaxation of audits as a signal to political supporters or as an indirect consequence of the slackness prevailing in public administration around elections.

Considering that a typical pre-election period has duration of circa 40 days, Fig. 8 compares the revenue in the two months of the election period in each electoral year with the same two months in adjacent years. Simple inspection shows that in most of the elections held between 1974 and 2009, average bimonthly revenues expressed as percent of GDP were lower than the average of the respective figures in the two adjacent years, (with



Greek Crisis in Perspective: Origins, Effects and Ways-Out, Fig. 8 Comparison of bimonthly tax revues in pre-election periods. Revenues are calculated for the period of two months before each election as % of annual GDP. Each election year (N) is denoted by black and compared with revenues collected over the same period during the previous (N − 2, N − 1) and the following years (N+1, N+2) denoted by grey. Frequency is bimonthly to account for the fact that the pre-election period lasts for 30–40 days, thus it extends over the prior as well as the poll month. Data are not seasonally adjusted, thus they reflect within year variations (Source: Skouras and Christodoulakis (2011) where further details are available)

only two slight exceptions in 2000 that coincided with the entry to EMU and 2007 because it is compared with another – and a lot worse – electoral period in 2009). In the same study it is estimated that pre-electoral misgovernance causes a loss in revenues equal to 0.18% of GDP in each election year. For the 13 elections taken place in the period 1974–2009, this amounts to more than 5 billion Euros at 2010 prices.

An Ex Post Assessment of the Memorandum

EU authorities seemed to be unprepared to react promptly and concertedly to the Greek problem and undertook action only when they recognized the risks it posed for the banking systems of other European states. After difficult negotiations, a joint loan of Euro 110 bn was finally agreed in May 2010 by the EU and the IMF to substitute for inaccessible market borrowing. The condition was that Greece follows a Memorandum of fiscal adjustments to stabilize the deficit and structural reforms to restore competitiveness. More details are given in the Appendix. In the eventuality of success, Greece would be ready to tap markets in 2012 and then follow a path of lowering deficits and higher growth. More than two years after implementation, the record remains poor and the economy is fiercely contracting. An explanation is attempted below.

The Failure in Fiscal Adjustment

The decline of revenues as a share to GDP after 2007 and the collapse of the collection mechanism in 2009 in particular were instrumental for the explosion of public debt and deficit thereafter. Strangely enough, no serious effort was undertaken to remedy the situation after the elections. The ministerial post in the Inland Revenue remained empty for more than a year and two top executives resigned in protest that their proposals to beat tax evasion were turned down. The Government opted for an increase in the VAT rate from 19% to 23% in the spring 2010 and, as a result, CPI inflation jumped to 4.5%, further cutting purchasing power amid recession. The only result was that activity was reduced and revenues did not rise.

The Government continued to act in a positive feedback loop, with lower revenues prompting higher taxation and this in turn causing further evasion. Unable to raise efficiency and under pressure to collect revenues, it imposed a heavy increase in fuel tax, substantial consumption surcharges and finally a lump-sum tax in exchange for settling previous arrears. Once again tax revenues ended up far below the target in a typical manifestation of elementary Laffer-curve predictions.

Only by the end of 2011 it was recognized that further tax measures are no longer viable and attention should shift on collection efficiency. In its assessment of progress, the European Commission task force warned that "... tax and expenditure measures ... substantially compress the households' disposable income and significantly tighten their liquidity constraints", (European Commission 2011, p. 2).

But that was no more than a void warning, because at the same time the Government was forced by the very same task force to retroactively raise the tax rate on the self-employed and impose a new levy on property in order to make up for the falling revenues.

Regarding public expenditure, a more optimistic picture emerged but at a huge cost in terms of growth and efficiency. Soon after the elections, the Government made clear signals that it had no real intention of containing the oversized public sector. Numerous appointments that were made before elections through a highly disputed process were nevertheless approved by the new incumbent, and a widely publicized operation to abolish and merge outdated public entities has made no real progress, to date. A novel scheme to push older staff onto a stand-by status with a fraction of their salary misfired as it was soon discovered that most on the list were exploiting the incentives of the system for an early retirement. After the fiasco the Government announced a lengthy process of evaluation in the public sector as a precondition for staff redundancies, but without setting a time limit it proved to be only an excuse to avoid actual decisionmaking.

In the absence of any structural adjustment in the public sector, the reduction of spending was achieved by imposing universal cuts in salaries and this led to widespread shirking practices. Another unusual tool for keeping expenditure low was to cut the budgetary co-financing of the European Community Support Framework, thus reducing public investment at a time when it was mostly needed to induce some growth in the economy. After the Decision by the European summit in July 2011, Greece was freed from the co-financing obligation, but when the new practice started to be implemented at the end of 2011 it was already too late to rectify the damage done to economic activity.

The Limits of Structural Adjustment

In order to rebalance the economy onto a more competitive path, the Memorandum agreement envisaged a long list of structural reforms ranging from reforming the social security system to removing closed-shop vocational practices, and from cutting red tape to liberalizing the licensing process for lorry and taxi drivers. The pension reforms initially succeeded in harnessing the deficits in the social security funds, but soon they reappeared when a wave of retirement took place in anticipation of imposing further age extensions in the future.

Most of the reforms were either abandoned or backfired. For example, the opening-up of lorry licenses failed to reduce transportation costs and enhance competitiveness in practice despite the severity of clashes with trade union hardliners. The reason for this was that insiders took advantage of a two year postponement and decided to maximize rent-seeking by withdrawing previous price concessions. Besides, the economic gloom was thwarting potential investors by making the upfront cost of setting up a new business too high.

A similar attempt to open-up the taxi licensing system was abandoned after a protracted clash with insiders in the summer 2011 that seriously damaged tourism in its period of peak. In other professions, such as lawyers and pharmacists, there was only a token liberalization without any reduction in consumer prices. Recognizing this failure, the new conditionality program imposed a regressive mechanism with the aim of reducing the overall profit margin to below 15%, (see Memorandum II 2012, para 2.8, "*Pricing of medicines*"). The results of this are still to be seen.

Seeing that the structural adjustment program was derailed, the Memorandum sought for alternatives. To enhance competitiveness in the labour market, liberalization measures extended part-time employment, imposed wage cuts across the board and removed collective bargaining agreements. Despite lowering labour costs by 12%, enterprises were overwhelmed by recession and unemployment became rampant, exceeding 17% of the labour force by the end of 2011. As with the positive feedback mechanism on the tax front, the rise in unemployment invited a new round of wage cuts in the private sector, shrinking further disposable income and fuelling new waves of social protest. Though the IMF mission in the autumn 2011 was explicit that "accelerated private sector adjustment . . . would likely lead to a downward spiral of fiscal austerity, falling incomes and depressed sentiment", it nevertheless urged for further such measures in order to achieve a ". . . critical of reforms needed to transform the investment climate", (IMF 2011). Bringing-up some growth to the real economy is still not a top priority for the program overseers.

The Failure of Privatizations

The failure of the privatization program is worth commenting on, as it reveals an unusual combination of strong rhetoric in theory with complete apathy in practice. Immediately after the elections in 2009, the Government showed that it had no intention of curbing the wider public sector. Its lack of resolve to tackle the excessive demands of public trade unions was made manifest in a dispute with a newly arrived investor in the Piraeus Port Company. The Government succumbed to paying enormous compensation for early retirement as a condition that the investment goes ahead. No privatization target was included in the 2010 Budget and none was actually implemented.

Thus it was viewed as a major shift of policy when the Government agreed in March 2011 to adopt a large-scale privatization plan of Euro

50 bn during the period 2011–2015, or roughly 4% of GDP per annum. The plan included extensive sales of public real-estate, privatizations of public enterprises in the energy sector and private partnerships in the operation of airports and ports throughout Greece.

After months of procrastination a market-friendly Privatization Fund was finally set up to replace the ineffectual authority that was in charge before, but its determination was this time hindered by adverse market conditions. With asset prices falling to abysmal levels, privatizations would be probably embarrassing in political terms and inadequate in terms of revenue, but in practice there was no real demand, as capital flight continued to be fuelled by fears of abandoning the Eurozone and funds from abroad were not coming for the same reason. Despite initial ambitions, the program achieved little in 2011, selling only an option on Greek Telecom, future rights to the National Lottery and publishing a preliminary tender for the redevelopment of the old Athens airport. In 2012 the program was downscaled to a meager Euro 2.8 bn, just a quarter of the amount initially announced.

The New Memorandum Conditionalities and Ways-Out of the Crisis

Faced with a deepening recession and a failure to produce fiscal surpluses sufficient to guarantee the sustainability of Greek debt, the European Union intervened twice to revise the terms of the Memorandum. In the first major intervention in July 2011, the amount of aid was increased substantially by Euro 130 bn and repayment was extended over a longer period of time. To implement the Private Sector Involvement (PSI) in debt restructuring, a cut of 21% of the nominal value of Greek bonds and re-profiling of maturities was decided upon with the tacit agreement of major European banks.

Crucially, the EU authorities this time fully recognized the perils of recession and allowed Greece to withdraw a total amount of Euro 17 bn from Structural Funds without applying the fiscal brake of national co-financing. The plan looked powerful, except for the typical implementation lags. The Agreement was only voted through by all member-state Parliaments only in late September 2011 and the release of structural funds was approved by European Parliament in late November. Participation in the PSI had reached only 70% of institutional holders amid speculation that post-agreement buyers of Greek debt from the heavily discounted secondary market were expecting a huge profit through their offer to cut it!

Thus, a new intervention looked inevitable and in October 2011 a revised restructuring (the so called PSI+) was authorized, envisaging cuts of 50% of nominal bond value that would eventually reduce Greek debt by Euro 120 bn and would allow it to be stabilized at around 120% of GDP by year 2020. In exchange, Greece would undertake further fiscal cuts of around 6% of GDP. Greek bonds held by public institutions (i.e. by the bail-out providers) will be fully honored, though social security funds and domestic public entities were forced to participate in the scheme.

The agreement was hailed as the definite solution to the debt conundrum, but euphoria turned sour a few days later when the Greek Government surprised everybody by seeking a referendum for its approval. Many feared that the outcome could in all probability be negative as an expression of current misgivings, and this would be quickly interpreted as opting for exiting the Eurozone. In the ensuing *furore*, the decision was annulled, the Prime Minister resigned and a coalition Government was formed in November 2011 to implement the restructuring of debt and negotiate the terms for the new round of EU-IMF loans. The Government acted quickly and concluded the PSI agreement in May 2012, but the extra fiscal package was not finalized as new elections were called for by the coalition partners, anxious to refresh their mandate before the political cost of further cuts becomes too inhibitive for them to remain in power. A caretaker Government followed and pre-electoral inaction adjourned privatizations so as not to excite opposition from the unions, and made collection authorities to slacken the processing of income tax statements so as not to infuriate the voters with an increased tax burden.

Revenues dropped significantly, sadly confirming the cycle described in Section 4.

The situation was further aggravated after the first election round proved inconclusive, with the two mainstream parties saw their share of the vote collapse from around 80% of the electorate in 2009 to just one third of the total in 2012, while leftwing and rightwing parties, with a strong rhetoric against the bail-out agreements, saw their share of the vote soar; this prompted a new wave of speculation that Greece is likely to exit the Eurozone and capital flight drained significant amounts from the Greek banking system.

The second round was polarized by the Euro dilemma, thus helping to increase mobilization of voters and finally resulting in a tripartite coalition that vowed to take all necessary measures to safeguard Greece in the Eurozone.

Though the new Government exhibits some of the weaknesses typical of party coalitions, it strives to persuade domestic and European opinion that it means business. As a signal, it reaffirmed that privatization of several public companies will go ahead and announced that the state will abolish minimum holding rights in utilities as an incentive to potential buyers. The Agricultural Bank was swiftly privatized and other state-owned credit organizations are to follow suit. The fiscal package is expected to be voted on in Autumn 2012, though a new round of social tensions and political breakaways cannot be ruled out.

In the meantime, two factors have adversely affected the situation of Greece versus the Eurozone and the bail-out authorities: On the European front, the market pressure is currently directed towards the economies of Spain and Italy, and the concomitant threat to the very existence of the Euro has pushed the Greek problem to the sidelines. Although European authorities responded to the challenge by designing a new defense mechanism in the banking system and the European Central Bank decided to intervene in the bond markets to stave off speculative attacks against member-states, European public opinion is characterized by a "*rescue fatigue*" and appears to be increasingly hostile towards granting any further support for Greece. However, Greece itself is overstressed: unemployment has rocketed above 24% of the labour force; recession has further deepened with the contraction of real GDP expected to exceed 7% in 2012, an awesome deterioration from milder estimates just a few months ago.

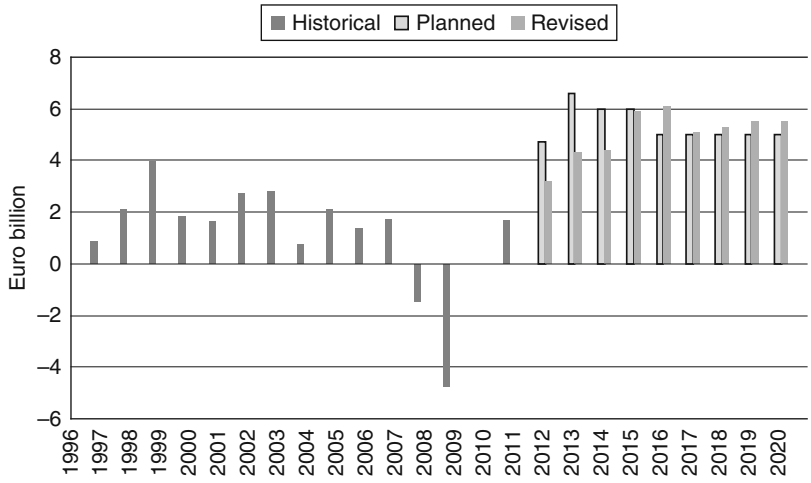
If this trend continues, debt stabilization will be jeopardized and a new cycle of futility and desperation will emerge. Since the dynamics of the debt-to-GDP ratio are sensitive to the prospects of growth, it is worth examining alternative debt paths that correspond to lower and higher growth profiles.

Three Alternative Scenarios: Walking on a Tight Rope

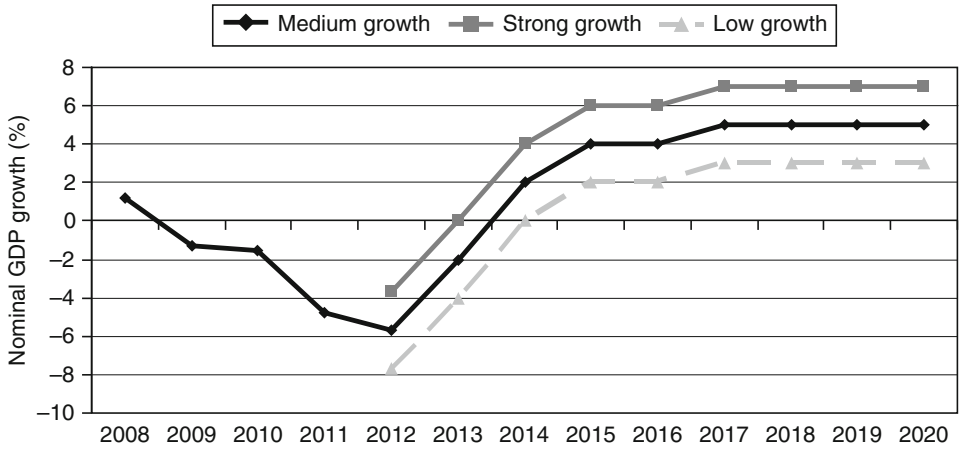
The alternatives revolve around a medium growth path, as has been calculated by European authorities in March (EC 2012, Table p 30) to ensure sustainability of public debt. This is based on real growth resuming in 2014 and staying above 2% thereafter, while inflation of GDP deflator is consistently kept below 2%. A primary budget surplus is achieved in 2013 and stays above 4% of GDP for the next of the period. Privatizations are to follow the revised schedule shown in Fig. 9, and the cost of borrowing is set at 4% per annum.

The lower and higher growth scenarios are devised by assuming two nominal growth paths cut or augmented by two percentage units respectively, as depicted in Fig. 10. All other assumptions remain the same as in the EC scenario to facilitate comparisons. Debt profiles are shown in Fig. 11.

The low-growth case leads to a debt to output ratio in 2020 above the level it had in 2010 when the country initially asked for a bailout. Most probably, the economy will collapse before the end of the period as generating primary budget surpluses of above 4% of GDP or collecting privatization proceeds of more than 2% of GDP for six consecutive years is utterly unrealistic under anemic growth. The low growth scenario is not out of context though, and, for a start, it replicates what actuall happened in 2012 with real growth rate plummeting at − 7% rather than the − 4.7% rate envisaged in the official scenario. If this gloomy trend continues, markets will pick up the



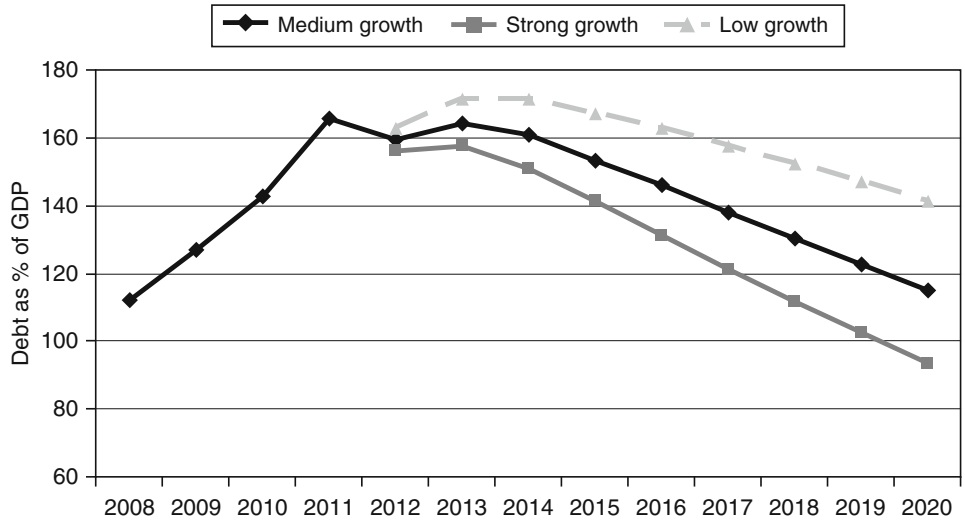
Greek Crisis in Perspective: Origins, Effects and Ways-Out, Fig. 9 Proceeds from privatizations. Note: Proceeds are net of capitalizations in state-owned enterprises. For 2008, 2009 and 2010 figures of proceeds are net of bank shares purchases, thus the negative sign (Source: Privatization Report, Ministry of Finance, 2008. Data for 1996 and 1997 are taken from Budget Reports. Planned figures were set in May 2011, but then they were revised in 2012) (Memorandum II, para 2.1)



Greek Crisis in Perspective: Origins, Effects and Ways-Out, Fig. 10 The three alternatives for recovery. Growth rates of nominal GDP are taken as the sum of real growth and the projected rates of GDP deflator. The medium growth rate is taken from European Economy (2012, Table p. 30), The strong and low growth profiles are obtained by simply assuming plus and minus two percentage units per year over the medium path

conundrum and the situation will soon get out of control. There is no political force in Greece eager to undertake new painful cuts on top of the current ones, and the country will be forced to abandon the stabilization program. European Governments – unable to ignore indignant public opinion – will insist on no more bailout aid without honoring previous obligations, and then Greece will be left impotent and unwilling to continue any further. End of game.

The higher growth scenario, on the other hand, leads to a debt to output constantly declining and reaching a level of around 95% at the end of the period, substantially below the medium scenario. The growth path is not unrealistic and real growth rates have just to be close to those that prevailed in



Greek Crisis in Perspective: Origins, Effects and Ways-Out, Fig. 11 Alternative paths for public debt as % of GDP. Note: Each scenario corresponds to an assumption of nominal growth in 2012–2020 shown in Fig. 10. All other assumptions remain the same. The medium growth scenario replicates EC (2012)

the previous decade, though now based on deep market reforms and without the fiscal extravaganza. Other assumptions, such as those of substantial primary surpluses and uninterrupted privatizations, also become more realistic with higher growth. Sustainability may be further assisted if the bail-out funds currently allocated for the recapitalization of Greek banks are taken out of public debt accounts and become liabilities of the new banking authority that is scheduled to operate next year on a Eurozone-wide basis. This will mean a further reduction of the debt to output ratio by more than twentyfive percentage units and will firmly anchor Greece in the Eurozone.

Is Exit from the Euro an Option?

The crisis in Greece had profound ramifications for the Eurozone, both in political as well as in economic terms. In the Euro area, Greece is routinely considered not only as devouring European taxpayers, but also as the habitual wrongdoer especially when compared with the other two countries (Ireland and Portugal) which are undergoing similar adjustment programs with more efficacy. In such a politically unyielding and increasingly suspicious framework, a Greek exit from the Eurozone started to attract attention both at home and abroad.

Though complications and costs that would ensue in the banking sector will be enormous, the exit of Greece could prove opportunistically attractive to some European politicians who get angrier every time a new round of aid is discussed. However, they overlook the fact that a Greek exit would reverberate around other states and lead to an aggravation of the crisis; for how contagion will spread see Vehrkamp (2011). It may also serve as the convenient argument for consolidating and enforcing a two-tier model of Economic Governance, as has been advocated before the creation of EMU (e.g. Bayoumi and Eichengreen 1992) and is recently suggested again by commentators and politicians betting on the "*Grexit scenario*" and assuming that other countries may follow suit. Based on an inner core of surplus economies in the north and a weaker periphery in the south, competitiveness in this model will be restored through the so called "*internal devaluation*" of labour costs, thus perpetuating the gap that is already widening between the Eurozone countries; for a description of divergences within the common currency see Christodoulakis (2009).

For Greece, exit would trigger a prolonged economic catastrophe. As the entire Greek debt will remain denominated in Euros, the rapid depreciation of the new national currency will

make its servicing unbearable and the next move will be a disorderly default. Isolation from international markets would drive investors even further away, while the financial panic would drain domestic liquidity at a massive scale. The creditor countries of the EU would start demanding repayment of their aid loans, and this would soon deprive Greece of its claim on the EU cohesion funds. Tensions are likely to produce further conflicts with EU agencies and the pressure to consider complete disengagement from the European Union will gain momentum both domestically and abroad.

Stay in the Eurozone and Grow More

The cost would be so immense that the single available option for Greece is to complete the fiscal adjustment and become reintegrated into the Eurozone as a normal partner. This requires Greece to undertake concrete actions that produce visible results within a short timeframe, so that society becomes more confident to pursue further reforms. Some policy suggestions for this direction are as follows:

First, Greece needs to acquire credibility while also being properly understood abroad. The continuing fiscal shortfall is easily translated as reluctance, causing continual friction with the European Union and demands for a new battery of austerity measures. To escape this cycle, Greece must adopt a front-loaded policy as a matter of urgency to achieve key fiscal targets quickly and to change the impression of being a tactical waverer. This seems to be the line adopted by the new Government. If Greece succeeds in this front-loaded policy, it may be in a position to revise some of the pressing – although so far unattainable – schedules and ensure greater social approval and tolerance. To ensure that there will be no spending spree in future elections, the best option for Greece is to adopt a constitutional amendment on debt and deficit ceilings, just as Spain did in 2011, alleviating market pressures, at least for the time being.

Second, Greece desperately needs a fast-track policy for exiting the long recession. An amount of Euro 17 billion could be disbursed and routed immediately to support major infrastructural projects and private investment in export-oriented companies. The growth-bazooka should then be followed by structural reforms and privatizations that can attract significant private investment as market sentiment is restored. In addition, instilling growth will help to control the debt dynamics and reduce public deficits without ever-rising taxes that thwart private investment and make economic recovery and sustainability even more unattainable. Feldstein (2012) leaves no doubt about the mechanics of stabilisation when he warns that "(t)o achieve a sustainable path, Greece must start reducing the ratio of its national debt to GDP. *This will be virtually impossible as long as Greece's real GDP is declining*", (my emphasis).

The inevitability of the above thesis cannot be ignored anymore. Nor can it be circumvented by sermons on the necessity of front-loaded reforms on the assumption that will automatically restore growth and competitiveness. In fact, the need for further growth spreads fast to other countries either in or outside the Eurozone; for example, the UK Government recently decided to inject BPS 50 bn on infrastructural projects to speed economic recovery and one just hopes that the Eurozone will be equally responsive to the need.

Conclusions

Exactly three decades after becoming a fully-fledged member of the European Union and ten years after joining the Eurozone, Greece sought a bail-out agreement in 2010 to avoid bankruptcy. A long history of stabilization programs proved incapable of achieving a lasting fiscal correction and adequately raising competitiveness, as fundamental weaknesses in the economic and political system continue to play a corrosive role. The oversized public sector and the frequent indulgence in pre-electoral spending sprees in exchange for political support led to protracted fiscal deficits and the accumulation of a large public debt. Equally, the chronic deterrence of productive investment by a multitude of regulatory inefficiencies resulted in a thin tradeable sector and large Current Account deficits. The economy remains vulnerable to political

developments which are often dictated by short-term partisan considerations with far reaching fiscal implications. This explains why, in spite of substantial reforms taking place over the last two decades and achieving high growth rates, EMU participation and moderate debt stabilization, the situation went once more out of control.

Regarding the current crisis, the article described how prolonged external and fiscal deficits were allowed to reach uncontrollable levels and, in the aftermath of the credit crunch, led to a further escalation of debt and the subsequent bail-out. Two and a half years later, fiscal consolidation is still far from being sustainable in spite of augmenting the bail-out loans and implementing a substantial debt reduction on private holders.

The economy has contracted by nearly 20% since 2008, social tensions are multiplying and the future of Greece in the Eurozone is in jeopardy. Some consider such an outcome as a due punishment for past excesses, while others see it as an escape from further unemployment and recession. The article finds both angles of view as illusory, and argues that the only viable way out of the current crisis is to restore growth and then adopt a realistic plan for privatizations and reforms. The lesson of the past two years is that deep recession will otherwise continue to hinder any existing possibility for exiting the crisis. Greece, and other Eurozone countries too, are desperate for a "*corridor of confidence*", to use Keynes' famous phrase, to put things in order before it is too late.

See Also

- ▶ European Central Bank
- ▶ European Central Bank and Monetary Policy in the Euro Area
- ▶ European Cohesion Policy
- ▶ European Monetary Integration
- ▶ European Monetary Union
- ▶ European Union Budget
- ▶ European Union (EU) Trade Policy
- ▶ Euro Zone Crisis 2010

Acknowledgments I have benefited from various comments by V. Sarantides and A. Ntantzopoulos and I am also thankful to participants in a LSE seminar on an earlier version of the paper.

Disclaimer Views expressed in this article are solely those of the author, without implicating or representing any other person or organization.

Appendix: A Brief Description of the Conditionality Programs for Greece

The adjustment program for Greece was laid out in three phases. The first Memorandum was signed in May 2010 and aimed at reducing the fiscal deficit to 3% in 2013. Specific measures that were actually implemented included universal cuts in public salaries and all pensions, a rise in VAT from 19% to 23% and similarly in other consumption surcharges, the abolition of collective agreements in favor of firm-level contracts, the lowering of private sector wages by 12% and a reform in the Social Security system. It also included the liberalization of red-tape practices in the transport sector, pharmacists and lawyers, but the outcome was heavily compromised through a series of delays and back offs. Fiscal deficit for 2010 ended up close to 11% of GDP, substantially lower than the horrendous 15.4% in the year before but still away from the initially set target.

Thus, in early 2011 a new round of negotiations resulted in a second round of measures voted by Parliament in June 2011. They included further taxation on past incomes, a lump-sum tax on professionals, further rises in indirect taxes and a new property levy that was imposed two months later. The program demanded the abolition of outdated public entities, the reduction in the number of civil servants and a further curtailment in their salaries. It also envisaged ambitious privatizations on utilities and public real-estate that could trim down public debt by Euro 50 bn within a four-year period. Fiscal deficit for 2011 is provisionally estimated to be 9.8% of GDP, revealing a major difficulty in further adjustment in the absence of growth.

The third round of adjustments was voted for in February 2012 as Memorandum II. (For the full text see "Memorandum of Understanding on Specific Economic Policy Conditionality", 9 February 2012, available at http://www.hellenicparliament.gr).

This time it was approved by the two major parties, but only after a line-up was imposed to avoid desertions and rising internal protest. Measures included a reduction of minimum wages in the private sector by 22%, an additional cut by 10% to new entrants as a means to beat youth unemployment, 15% cuts in various pensions, the abolition of several tax credits and explicit targets for cutting employment and entities in the wider public sector. Policies will start to be implemented in the final quarter of 2012.

Bibliography

Bayoumi, T., and Barry, Eichengreen. 1992. Shocking aspects of European monetary unification. CEPR discussion paper no. 643, May.

Blanchard, O. 2006. Current account deficits in rich countries. *IMF* Mundell-Fleming Lecture.

Blanchard, O., and F. Giavazzi. 2002. Current account deficits in the Euro area: The end of the Feldstein-Horioka puzzle? *Brookings Papers on Economic Activity* 33: 147–210.

Christodoulakis, N. 1994. Fiscal developments in Greece 1980–93: A critical review. *European Economy: Towards Greater Fiscal Discipline* 3: 97–134 . Brussels.

Christodoulakis, N. 2009. Ten years of EMU: Convergence, divergence and new priorities. *National Institute Economic Review* 208: 86–100 .London

Christodoulakis, N. 2010. Crisis, threats and ways out for the Greek economy. *Cyprus Economic Policy Review* 4(1): 89–96.

Christodoulakis, N. 2012. Market reforms in Greece 1990–2008: Domestic limitations and external discipline. In *Market reforms in Greece*, ed. Kalyvas, S. and G. Pagoulatos. Columbia University Press (forthcoming).

Christodoulakis, N., and V. Sarantides. 2011. External asymmetries in the Euro area and the role of foreign direct investment, *Bank of Greece.* Discussion paper.

De Grauwe, Paul. 2010. The Greek crisis and the future of the eurozone. *Intereconomics* 2: 89–93.

European Commission. 2009. Quarterly report on the Euro area 8(1), Brussels.

European Economy. 2011. The economic adjustment programme for Greece: Fifth review. Occasional papers 87, October.

European Economy. 2012. The second adjustment programme for Greece: Fifth review. Occasional papers 94, March.

Feldstein, Martin. 2012. The failure of the Euro: The little currency that couldn't. *Foreign Affairs* 91(1): 105–116.

IMF. 2011. Greece: Third review under the stand-by arrangement. International Monetary Fund, Country Report No. 11/68, March.

Krugman, Paul. 2011. Origins of the euro crisis, *blog* "The Conscience of a Liberal," 23 September.

Memorandum, II. 2012. Memorandum of understanding on specific economic policy conditionality, 9 February 2012. Available at http://www.hellenicparliament.gr

Shelburne, R.C. 2008. Current account deficits in European emerging markets. UN discussion paper, no. 2008. 2.

Skouras, S., and Christodoulakis, N. 2011. Electoral misgovernance cycles: Evidence from wildfires and tax evasion in Greece and elsewhere. LSE, Hellenic Observatory. GreeSE paper no. 47.

Vehrkamp, R. 2011. Who's next? The Eurozone in an insolvency trap. *Bertelsmann Stiftung*, no. 2.

Green National Accounting

Sjak Smulders

Abstract

Extending conventional national product measures, green national accounting provides better indicators of economic welfare and of the sustainability of welfare levels. The main theoretical result shows that in an undistorted economy net national product is proportional to welfare, provided some rather stringent conditions are met. With appropriately used shadow prices, the welfare effects of externalities and world market changes can be accounted for and sustainable income – the hypothetical level of consumption that can be sustained into the future – can be calculated. Practical approaches have been proposed to adjust conventional national income figures roughly in the spirit of the theoretical results.

Keywords

Air quality; Climate change; Contingent valuation; Ecological footprint; Education; Environment; Genuine savings; Green national accounting; Green net investment; Green net national product; Hedonic prices; Household work; Index of sustainable economic welfare; Inequality of income; Intergenerational income distribution; Interpersonal utility comparisons; Land use; Leisure; Maximin preferences; National accounting; Natural resources; Non-renewable resources; Opportunity cost; Other-regarding preferences; Pollution; Representative agent; Resource depletion; Shadow prices; Social welfare function; Sustainability; Sustainability accounting; Sustainability gap; Technical change; Travel costs; Utilitarian preferences; Utility aggregation; Value of time; Wealth

JEL Classifications

O47; Q01

Green national accounting extends conventional national product measures to provide better indicators of economic welfare, as well as indicators of the degree to which welfare levels can be sustained.

Conventional national accounts measure the size of the market or commercial activities, but do not necessarily measure very well (a) how these activities translate into welfare and (b) how non-marketed activities goods contribute to the welfare of citizens. A big part of the greening of national accounts concerns issues related to the environment. For example, production of certain goods that generate market value contributes to national income, but, if the production generates undesirable pollution as a by-product, the contribution to welfare might be actually negative. Conventional national accounts ignore the reduction in air quality since air quality is not traded on markets and is left out from conventional national accounts. As another example, consider the depletion of oil reserves. Oil companies' profits contribute to conventional national income, but the fact that fewer reserves will be available in future is left unaccounted, even though this may affect the welfare of future generations.

The aim of measuring economic welfare by a simple number is without doubt challenging, since it is inevitably related to the formal economic concept of individual utility and the theoretical problems of aggregating utility and interpersonal utility comparisons. The theoretical approaches to green accounting circumvent these problems by assuming a social welfare function and representative agents in highly stylized models. Because of lack of data and various other constraints, the more applied approaches to green accounting are only loosely rooted in formal theory and sometimes include issues of (intergenerational) income distribution on an ad hoc basis.

The focus of green accounting is much more on dynamic and intergenerational aspects than on intra-generational issues. It typically tries to measure to what degree the lifetime utility of the representative agent in a country increases over time, and to what degree it is higher than in other countries. It sometimes also tries to measure to what degree intra-temporal levels of utility of the representative agent in a country can be maintained over time, whether the economy is investing enough to maintain non-decreasing utility levels, and how much a country can consume more than another country when both of them would ensure utility levels of their inhabitants are not declining over time. The latter type of questions is often associated with 'sustainability accounting' as a particular branch of green accounting.

Green accounting starts from a broad concept of economic welfare, which goes beyond welfare depending on just marketed produced goods. Thus, welfare is allowed to depend on health, environmental amenities, pollution levels, or availability of natural resources. Even altruistic preferences are allowed: utility levels of future generations may matter for the welfare objective of current generations. Society might care in particular for the utility levels of those generations that are worse off in future. The extreme case of this implies *maximin preferences:* only the generation with lowest utility levels gets a positive

weight in the social welfare function and reductions in other generations' welfare do not count. This contrasts to *utilitarian preferences*, in which every generation gets a weight, and utility levels of generations further in future often get a lower weight because of a positive utility discount rate.

Welfare in an Undistorted Economy

A fundamental theoretical result concerns the measurement in undistorted competitive (and hence by construction welfare-maximizing) economies (Weitzman 1976, 2003), which we will refer to as the *Weitzman principle.* Consider a society that manages to maximize its own social welfare function, either because there are no externalities or because all existing externalities are internalized by appropriate policy. In such an economy, green net national product (NNP) can be calculated and this measure is proportional to total welfare. Moreover, green net investment can be calculated and a positive (negative) value of this measure always implies an increase (decrease) in instantaneous welfare.

To start with, we consider the simplest model economy (as in Weitzman 1976), in which a single consumption good is produced from a single capital good, representative agents maximize utility discounted at a constant rate, and the social welfare function is the sum of individual utilities. Then the sum of the value of consumption and net investment, valued against market prices, is proportional to welfare. Note that this sum is equal to the conventional NNP number for this economy. In this model economy, NNP reflects intertemporal welfare: consumption reflects instantaneous utility, whereas investment reflects how current economic activity contributes to future utility.

Extending conventional income to 'green' income is needed if the welfare function has 'unconventional' arguments like environmental quality and health (see, for example, Asheim and Buchholz 2004). These arguments can be seen as alternative forms of consumption, not consumption of conventionally produced goods but of natural resource services, health services, and so on. Then, in a perfect economy, according to the Weitzman principle, green NNP should be calculated as the value of all 'consumption' activities that matter for utility plus the value of net investment in all 'capital' stocks that matter for production capacity. Both 'consumption' and 'capital' are broad comprehensive measures here, with the former including for example consumption of environmental resource services and the latter including any variable that determines the production capacity for the generation of comprehensive consumption. Accordingly, the relevant capital goods, or assets, include not only physical capital and resource stocks, but also all kind of other state variables like public health (in economies that have a preference for health or in which health determines workers productivity), atmospheric pollution stocks, physical characteristics of the soil determining absorption of pollution and regeneration of nature, institutional capital and social norms subject to erosion and development.

As a special result of the Weitzman principle, the *Hartwick rule* can be derived (Hartwick 1977). The rule says that, if society wants to maintain a constant utility level over time, it has to invest the returns to non-renewable resources in other assets such that total green net investment (the comprehensive measure of investment) is zero. Hence, in accordance with the Weitzman principle, in such an economy green NNP equals the sustainable level of utility (as well as green consumption, since green net investment is zero). As a result it is a measure of sustainability: comparing two different economies that are both undistorted and maximize maximin preferences, we can say that the country with higher green NNP can maintain indefinitely a higher welfare level than the other.

Caveats

The above results must be interpreted with care and are more limited than might seem at first sight. The important caveat is that an economy with zero green net investment is not necessarily able to maintain a constant utility level for its representative agent over time and can thus be unsustainable

(Asheim et al. 2003). This may be the case, for example, in an economy that is dependent on a non-renewable resource and that is maximizing a utilitarian welfare function with constant discount rate rather than a maximin welfare function. Such an economy might *optimally* consume growing amounts initially, but eventually consume declining amounts, which of course is always inefficient for maximin preferences. The key insight from the Hartwick rule is that a necessary, but not sufficient, condition for maintaining constant welfare over time is sufficient investment (Pezzey 2004).

The Weitzman principle applies only when changes in production capacity depend solely on investment choices. Alternatively, production capacity, that is, the possibility to generate consumption, might change over time due to events beyond control of the economy. Three important examples are exogenous technological change, world price changes, and geological or climatic changes. If technology improves over time or the world market price of export goods increases, an economy's welfare can improve even when green net investment (as defined above) is negative. Formally, welfare is now proportional to the sum of comprehensive consumption and net investment, augmented with a term capturing the (properly valued and discounted) benefits from improved technology and world market prices that will accrue to the economy in future (Sefton and Weale 1996). The latter term can be labelled the 'value of time', whereas the sum of green NNP and the value of time is 'augmented NNP' (Pezzey and Toman 2002). In the case of global warming or other negative environmental developments because of purely geological or climatic reasons, there might be a negative time premium and sufficiently positive green net investment is needed to keep welfare constant.

Externalities and Sustainable Income

The Weitzman principle is derived for welfare-maximizing economies. How can we measure welfare in economies that do not actually maximize welfare? Similarly, how can we measure sustainable income (consumption) levels in economies that do not actually sustain constant consumption levels?

One theoretically possible way is to construct a hypothetical income figure that represents the level of welfare or sustainable consumption, respectively, that would arise if the economy made the switch from being distorted to being welfare maximizing or sustainable, respectively. This requires a measurement of the total *wealth* of the economy, consisting of all assets valued at their corresponding shadow prices, which depend on the exact welfare function.

The problem when putting this into practice is that actual prices observed in the distorted economy have no direct relation to the shadow prices needed to calculate wealth. In the presence of externalities, market prices do not reflect certain social costs and benefits, so that the sum of consumption and net investment at market value misses some contributions to welfare. As an alternative to calculating wealth against shadow prices as indicated above, one may augment the market value of comprehensive consumption and net investment with the net present value of the marginal externality along the competitive path (Aronsson et al. 2004). Obviously, such an augmentation term in green accounting is also hard to calculate in practice.

Green National Accounting in Practice

Given these theoretical results, how feasible is welfare and sustainability measurement in practice? The results suggest that we need (*a*) comprehensive accounting of consumption and investment activities, (*b*) the right (shadow) prices, and (*c*) additional forward-looking augmentation terms to capture exogenous or uninternalized developments over time. It has been concluded that (b) and (c) are insurmountable impediments to practical green accounting: fully correct green accounting is impossible and any method ignoring the problems associated with (*b*) and (c) is bound to produce biased numbers. At the other side of the debate it has been argued that national accounting has always been imperfect and indicative only (Cairns 2002). According

to this view the task is to focus on making national accounts more comprehensive – and satisfy at least requirement (*a*) – carefully delineating consumption and net investment, and applying corrections to prices where reasonable and feasible. Furthermore, the value of non-marketed activities has to be imputed rather than observed. A host of methods is available to impute prices, which use hedonic pricing, contingent valuation methods, and travel cost approaches.

The resulting practical approaches differ with respect to what to include in national accounts and how to determine values and prices. Among them 'genuine savings' (GS), a comprehensive net investment measure, is the best-known and closest to theory (Pearce and Atkinson 1993). Most GS correct conventional measures of investment for consumption of resources, damages from pollution, and investment in education. GS is often found to be positive in Europe and Japan (thanks to high savings and investment in education) and negative for Africa and oil- producing countries (due to the depletion of oil reserves). The latter results are quite sensitive to the way resource depletion is accounted for.

GS figures should be interpreted with care: since the GS calculation ignores 'value of time' terms and uses market prices rather than shadow price, GS is not a true measure of welfare increases. Persistent negative rates of GS are likely to result in decreases in welfare (unless exogenous technological change is significant), but with positive rates of GS nothing definitive can be said. So GS can be used to measure only unsustainability.

Another well-known indicator is the index of sustainable economic welfare (ISEW, initiated by Daly and Cobb 1989). It is an extended measure of green NNP (and therefore aimed at measuring welfare) that starts from conventional income, adds changes in environmental quality, imputes value of non-marketed activities (particularly household work), subtracts consumption expenditures that do not directly contribute to welfare (such as health and pollution abatement expenditures), and weighs on an ad hoc basis remaining expenditures by a measure of income inequality. For different components different valuation methods are used so that consistency is not always guaranteed. Shadow prices or opportunity costs are rarely used to value damages. The large number of adjustment to NNP also raises the question why certain components are still omitted (for example, the value of leisure time is not included while household work is, and investment in education and technological change are omitted).

Calculations of ISEW for richer countries show that ISEW has grown considerable more slowly than conventional NNP over recent decades. This result is not robust, however, for changes in the specific composition of the index and the valuation assumptions (Neumayer 2003).

Because of the main problem of determining correct prices for non-marketed goods and for dealing with externalities, but also because of the limited substitution between natural resources and conventional man-made inputs, it is often argued that purely physical indicators are useful to measure the state of the environment and economic welfare and to supplement (rather than adjust) more conventional measures. Since no attempts are made to monetize and the common denominator is missing, different physical indicators cannot be easily aggregated to overall welfare: improvement in one indicator cannot be compared with gains elsewhere and the costs of securing improvements cannot be determined.

The *sustainability gap* indicator is an example (Ekins and Simon 1999). Since welfare is critical dependent on air quality, a minimum level of air quality can be defined that is needed to maintain welfare at a reasonable level, and it can be measured how far society is from this standard; the exercise can be repeated for other critical natural resources.

Another popular example is 'ecological footprint', which measures the amount of land that is needed to generate the consumption of a country, including the land needed to assimilate the waste generated and undo climatic change from carbon dioxide emissions by means of carbon sequestration (Wackernagel and Rees 1996). If for the world as a whole the ecological footprint exceeds available land, the economy is said to be unsustainable. A problem here is how to aggregate over different land uses. Similar measures,

with a similar aggregation problem, keep track of varieties of material resource flows.

It is unlikely that the theory of green accounting can in the end be fully applied. Instead, a combination of different indicators and imperfect theory-based measures of welfare could – together with the caveats from theory – provide a useful information system to put conventional national income systems into proper perspective.

See Also

- ▶ Externalities
- ▶ Intertemporal Equilibrium and Efficiency
- ▶ Shadow Pricing
- ▶ Social Discount Rate
- ▶ Sustainability

Bibliography

Aronsson, T., K. Löfgren, and K. Backlund. 2004. *Welfare measurement in imperfect markets: A growth theoretical approach*. Cheltenham and Northampton: Edward Elgar Publishing.

Asheim, G., and W. Buchholz. 2004. A general approach to welfare measurement through national income accounting. *Scandinavian Journal of Economics* 106: 361–384.

Asheim, G., W. Buchholz, and C. Withagen. 2003. The Hartwick rule: myths and facts. *Environmental and Resource Economics* 25: 129–150.

Cairns, R.D. 2002. Green accounting using imperfect, current prices. *Environment and Development Economics* 7: 207–214.

Daly, H., and J. Cobb. 1989. *For the common good: Redirecting the economy toward community, the environment and a sustainable future*. Boston: Beacon.

Ekins, P., and S. Simon. 1999. The sustainability gap: A practical indicator of sustainability in the framework of the national income accounts. *International Journal of Sustainable Development* 2: 24–58.

Hartwick, J. 1977. Intergenerational equity and investing rents from exhaustible resources. *American Economic Review* 67: 972–974.

Neumayer, E. 2003. Weak versus strong sustainability: Exploring the limits of two opposing paradigms, 2nd rev. edn. Cheltenham and Northampton: Edward Elgar Publishing.

Pearce, D., and G. Atkinson. 1993. Capital theory and the measurement of sustainable development: An indicator of 'weak' sustainability. *Ecological Economics* 8: 103–108.

Pezzey, J. 2004. One-sided sustainability tests with amenities, and changes in technology, trade and population. *Journal of Environmental Economics and Management* 48: 613–631.

Pezzey, J., and M. Toman. 2002. Progress and problems in the economics of sustainability. In *The international yearbook of environmental and resource economics 2002/2003*, ed. T. Tietenberg and H. Folmer. Aldershot: Edward Elgar Publishing.

Sefton, J., and M. Weale. 1996. The net national product and exhaustible resources: The effects of foreign trade. *Journal of Public Economics* 61: 21–47.

Wackernagel, M., and W. Rees. 1996. *Our ecological footprint: Reducing human impact on the Earth*. Gabriola Island: New Society Publishers.

Weitzman, M. 1976. On the welfare significance of national product in a dynamic economy. *Quarterly Journal of Economics* 90: 156–162.

Weitzman, M. 2003. *Income, capital and the maximum principle*. Cambridge: Harvard University Press.

Gregory, Theodore Emanuel Gugenheim (1890–1970)

R. J. Bigg

Gregory was born in London and after an education at St Owen's School in Islington, at Stuttgart and at the London School of Economics, he became an assistant lecturer at the LSE from 1913 to 1919. Appointed Cassel Reader in International Trade in 1920, he became Dean of the Faculty of Economics in London (1927–1930) and was Sir Ernest Cassel Professor of Economics from 1927 to 1937.

Robbins refers to him as 'one of the last of the generation of gifted teachers who, in the twenties, contributed so much to the international standing of the London School of Economics' (*The Times*, 3 February 1970).

Gregory sat on various official commissions including the Macmillan Committee on Industry and Finance (1929–31), the Irish Free State Banking Commission (1934–37), and acted as economic adviser to the Indian government from 1938 to 1946. He was knighted in 1942.

Gregory was 'acutely conscious ... of the disintegrating elements at large in the world' in the interwar period (Robbins 1970). His interwar books on currency reflect this outlook, supporting a return to gold as less liable to inflationary abuse, despite some flaws, rather than a managed system. His analysis was meticulous, examining the problems of timing, parities and the long-term supply and demand outlook for gold. He argued that exchange rate stability reflects the stability of relative prices between countries; thus the best way of achieving this was to link prices to the same standard, i.e. gold. He argued strongly against the inflation of paper money, and although gold itself has no intrinsic stability of value he argued that experience had shown it to be more stable than managed currencies. He further suggested that the improved stability of some paper currencies was due to exchange rate targets based on the US dollar which had remained tied to gold (an effective Gold Exchange Standard).

His most significant work was concerned with the history of banking and currency culminating in his introduction to Tooke and Newmarch's *History of Prices* and a history of the Westminster Bank that was a valuable contribution to the study of nineteenth-century monetary history. However, Gregory's interests were not restricted to the monetary field and Schumpeter, alongside his commendation of the introduction to Tooke, praises Gregory's article on 'The Economics of Employment in England, 1660–1713' (*Economica* 1921) and says 'There is no equally valuable survey for any other country' (p. 272).

Selected Works

1921a. *Foreign exchange before, during and after the war*. World of Today Series. Oxford: Oxford University Press.
1921b. The economics of employment in England, 1660–1713. *Economica* 1: 37–51.
1925. *The return to gold*. London: Ernest Benn.
1928. *Introduction to Tooke and Newmarch's history of prices*. London: P.S. King & Son. Reprinted, London: LSE, 1962.
1929. *Select statutes, documents and reports relating to British banking 1832–1928*, 2 vols. London: Oxford University Press.
1936. *The westminster bank through a century*, 2 vols. London: Oxford University Press.

Gresham, Thomas (c1519–1579)

Eleanor G. Powell

Keywords
Gresham, T.

JEL Classifications
B31

The second son of Sir Richard Gresham, merchant, Sir Thomas Gresham was educated at Gonville Hall, Cambridge, apprenticed to his uncle Sir John Gresham, also a merchant, and admitted a member of the Mercers' company in 1543. In 1551 or 1552 he became royal agent or king's factor at Antwerp, in which post he received 20 shillings a day, and which he retained with few intervals during three reigns until 1574, employed in spite of his Protestant views even by Mary. His business was to negotiate royal loans with Flemish merchants, to buy arms and military stores, and to smuggle into England as much bullion as possible. He succeeded in raising the rate of exchange from 16 s. to 22 s. in the £, and is said to have saved in this way 100,000 marks to the crown and 300,000 to the nation. His operations greatly benefited English trade and credit, though the government could not be induced to pay its debts as punctually as Gresham would have liked. He did not hesitate to remonstrate with and advise Elizabeth and Cecil; but he was so useful and trustworthy that he was never seriously out of

favour, except just after Mary's accession. On Mary's death he advised Elizabeth to restore the base money, to contract little foreign debt, and to keep up her credit, especially with English merchants. Later he taught her how to make use of these English merchants when political troubles in the Netherlands curtailed her foreign resources; at his suggestion the Merchant Adventurers and Staplers were forced by detention of their fleets to advance money to the state; but as they obtained interest at 12 per cent instead of the legal maximum of 10, and the interest no longer went abroad, the transaction proved advantageous to all parties and increased Gresham's favour. His journeys to and from Antwerp were very frequent, but in his later years he entrusted most of his public work to his agent, and is not known to have been at Antwerp after 1567. In 1554 he was sent to Spain to procure bullion, a very difficult task in which he was only partially successful; and in 1559 he was employed as ambassador to the Duchess of Parma, regent of the Netherlands; it was on this occasion that he was knighted.

In addition to his public services he continued throughout his life to do the work of 'the greatest merchant in London'. He was, in the language of the day, a banker and goldsmith, with a shop in Lombard Street, as well as a mercer; but he was a considerable country gentleman besides, with estates, chiefly in Norfolk, where his father held considerable property, and with several country houses besides the house in Bishopsgate which he built and bequeathed to London as Gresham College. He twice entertained Queen Elizabeth as his guest. His wealth was mainly earned by his private business, but he cannot be acquitted of enriching himself at the public expense by at least one dishonourable manœuvre; and he habitually forwarded his schemes by bribery. The money so gained he applied to public uses, his only son having died young: the foundation of the royal exchange, of Gresham College, and of eight almshouses, and the establishment of the earliest English paper-mills on his estate at Osterley, show the breadth of his interests, his liberality, his charity, his culture, and his commercial enterprise.

Bibliography

Anon. 1883. *A brief memoir of Sir Thomas Gresham, with an abstract of his will, and of the act of parliament for the foundation of Gresham's College*. London.

Bindoff, S.T. 1973. *The fame of Sir Thomas Gresham*. London: Jonathan Cape.

Boorstin, D.J. 1980. *Gresham's law: Knowledge or information?* Washington, DC: Library of Congress.

Burgon, J.M. 1839. *Life and times of Sir Thomas Gresham*, 2 vols. London.

de Roover, R.A. 1949. *Gresham on foreign exchange*. Cambridge, MA: Harvard University Press.

Fox Bourne, H.R. 1866. *English merchants*. London.

Fuller, T. 1662. *Worthies of England*. London.

Hall, H. 1886. *Society in the Elizabethan age*. London.

Holinshed, R. 1807–8. *Chronicles of England, Scotland and Ireland,* 6 vols. London.

Salter, R.F. 1925. *Sir Thomas Gresham*. London: Parsons.

Ward, J. 1740. *Life of Sir Thomas Gresham*. London.

Gresham's Law

François R. Velde

Abstract

This article recounts the origin of Gresham's Law, discusses its theoretical and empirical problems, and presents several refinements that have been proposed.

Keywords

Assignats; Asymmetric information; Bimetallism; Exchange rate indeterminacy; Good and bad money; Giffen, R.; Gresham, T.; Gresham's law; Legal tender; Seigniorage

JEL Classifications

E4

Gresham's Law, which holds that 'bad money drives out good money', is as problematic as it is well-known. It predicts that, when two monies are in use but one is of lower quality or lower intrinsic value than the other, the former will be used as

medium of exchange to the exclusion of the latter. The law implicitly relies on the monies circulating for the same value in spite of their intrinsic differences. The key question is: why would they?

Origin of Gresham's Law

Gresham's Law is one of the 'laws' bequeathed to us by 19th-century political economists eager to uncover laws of nature just like physicists. Henry Dunning Macleod (1855–1856, vol. 2, p. xxxvi; 1858, p. 477; 1896, p. 38), the man who named it, described it successively 'an unerring law of nature', 'a fundamental and universal law in Economies, which has been found to be true in all countries and ages', and 'this great Law, which is as well and firmly established as the Law of Gravitation'.

Perhaps it should have been named Macaulay's Law, for it was Lord Macaulay (1850–1861, vol. 4, p. 620) who gave the law its familiar form: 'where good money and bad money are thrown into circulation together, the bad money drives out the good money.' Macaulay also noted that the phenomenon had been described in the 5th century BC by the playwright Aristophanes, who in *The Frogs* (718–733) compared the prevalence of bad politicians with the replacement in use of high- quality gold and silver coinage by inferior copper coins. Macaulay dismissed the playwright's explanation based on preferences but thought the verses worth quoting in the original Greek. MacLeod quoted Aristophanes, too, albeit in English, but he decided to credit the first person he thought to have explained the law properly: Sir Thomas Gresham, a financier in 16th-century England. Gresham's letter to Queen Elizabeth I on her accession, urging her to restore a good quality coinage after the debasements of her predecessors, had recently been published by Burgon (1839, vol. 1, p. 483).

Wolowski (1864, p. lix) soon drew attention to earlier formulations of the law in monetary tracts of Nicholas Copernicus (written between 1519 and 1528) and Nicole Oresme (written between 1355 and 1357), although the relevant passage in Oresme's tract was later shown to be an anonymous addition of the late 15th century (Bridrey 1906, pp. 263–5). Later, Fetter (1932) pointed out that Gresham never stated anything remotely approaching his law: the famous letter to Queen Elizabeth I merely stated that the price of the English pound sterling abroad had declined after Henry VIII's debasement of the silver currency and that gold had been exported on that occasion. The somewhat arid search for the first person to state the law is often led astray towards descriptions of the phenomenon in some particular instance (such as the aftermath of a debasement), or of a related phenomenon (such as clipping, culling, and exporting of good coins). On this score, Copernicus's claim of priority is solid, since he wrote as a general proposition that 'introducing a new, worse money while the old, better one remains not only depreciates the latter, but, I would say, expels it' (Wolowski 1864, p. 56).

Be that as it may, MacLeod's coinage popularized by Jevons (1875) soon gained universal currency and, being taken at face value, this 'law' has been driving critical thinking out of circulation ever since.

The Problems with Gresham's Law

In the usual laconic formulation of the law, 'bad money drives out good money', every single word cries out for clarification. What is 'money', what is 'good', what is 'bad', and what does 'driving out' mean? The nature of the law, empirical regularity or theoretical proposition, is just as uncertain.

The law, being so vague, has been applied to a wide variety of pairs of monies: freshly minted and worn or full-bodied and clipped versions of a given coin; original and debased version of a given coin; a coin and a similar but lighter coin of the same metal; coins of different metals; metallic and paper money; paper monies of different issuers; securities of varying risk characteristics; and so on. The process of 'driving out' is taken to mean that one money replaces the other in some monetary function either completely or

partially, but also that the coins circulate alongside but not at par (Fetter 1932). In formal models, Gresham's Law is often invoked to compare outcomes across equilibria (for some parameter values, two monies circulate, and for others only the bad one does) rather than the dynamic process suggested by the words 'drive out'.

As a theoretical proposition, the law seems to run against basic economic intuition. Bad apples do not drive out good apples, they fetch a lower price. Why the good money could not circulate at a premium is a puzzle that the mere invocation of the law occults. Friedman and Schwartz (1963, p. 27 n.) note that the law is often misused because it 'applies only when there is a fixed rate of exchange between the two [monies]'. This raises the question: what fixes the exchange rate? In this version, the law relies on a postulate about prices, which economists are in the business of explaining.

As an empirical regularity, the law would have been more compelling if counter-examples had not been ignored as carefully as examples have been collected. One need only go to the original inventors of the law to find instances of selective empirics. Lord Macaulay invoked it as a general proposition to explain the state of the English coinage in 1695, a time when high-quality milled money was driven out by worn and clipped hammered money. He does not explain, however, why the law failed to apply in the previous 30 years throughout which milled and hammered coins coexisted (see Sargent and Velde 2002, chapter 16). Sir Thomas Gresham himself was baffled by the fact that, after Henry VIII's debasements, testoons containing 40 grains of silver circulated alongside testoons containing 20 grains, and at the same value (De Roover 1949, p. 93). Other counter-examples are easily found. High-quality currencies have dominated lower-quality competitors in international trade since the days of the Florentine florin and the Venetian ducat. Rolnick and Weber (1986) have documented numerous other violations of the law. Finally, the vast literature documenting collapses in real balances during hyperinflations surely testifies to the fact that (very) bad money will be driven out by almost anything else.

Refinements of Gresham's Law

Several refinements of the law have been proposed, and three are considered here.

One force that might fix the exchange rate is legal-tender laws. A law conferring legal tender status on a money states that a debtor is discharged of his debt by tendering that money (in the correct amount) to the creditor. It is up to the creditor to accept or refuse, but should he refuse he has no further legal recourse against the debtor. The argument that legal-tender laws can be sufficient to uphold Gresham's Law runs as follows. Suppose that two currencies, one intrinsically worth less than the other, are given equal legal tender by enforceable laws: say, one dollar coin is worth 90% of the other in intrinsic content. Debtors will then discharge their debts with the bad money, and reserve the good money for other purposes. The bad money displaces the good one, at least in repayment of debts.

A first difficulty is that legal-tender laws typically apply to the discharge of debts, not to circulation in general. But – if we set this aside – while the law forces the creditor to take the bad dollar in payment, it usually does not force the debtor to tender only bad dollars, or the creditor to accept good dollars at face value. The debtor owing $100 could offer the creditor a choice between 100 bad dollars and 90 good dollars (on the assumption that the transactions costs of such an offer are not too high, as they might be if the relative price is an inconvenient fraction; see Rolnick and Weber 1986).

To shore up Gresham's Law requires unusually strong legal tender laws, such as the two examples given by Selgin (2003). In the first example, on 11 January 1776 the Continental Congress resolved to protect the paper money it was issuing, the continental, by declaring that

> if any person shall hereafter be so lost to all virtue and regard for his country, as to refuse to receive said bills in payment, or obstruct or discourage the currency or circulation thereof, and shall be duly convicted [...] such person shall be deemed, published, and treated as an enemy of his country, and precluded from all trade or intercourse with the inhabitants of these colonies.

In the second example, the French revolutionary government defended its paper currency, the assignat, by decreeing on 11 April 1793 that trading specie for the assignat at anything but par, and setting or offering different prices for payment in assignats or specie, would be punished by 6 years of imprisonment. A law of 5 September 1793 added that such acts committed with counter-revolutionary intent were punishable by death.

Such examples, however, are too infrequent to sustain Gresham's Law with any kind of generality. They also tend to occur in periods of turmoil when the force of the legal-tender laws is questionable. The Continental Congress's resolution, which Nussbaum (1950, p. 567) finds 'of dubious legal significance', did not prevent the continental from quickly depreciating below par. In the French case, the assignat was already trading at less than half the price of gold coins, according to the Treasury's own records, and never traded above thereafter. These strong laws succeeded in propping up the demand for these paper currencies, but certainly not at anything like par.

A second refinement was proposed by Walker (1877, pp. 193–5) following Ricardo, and was endorsed by Giffen (1891) and by Palgrave (1894–1899, vol. 2, p. 262). Bad money will drive out good money only when the sum of the two is in excess of the wants of trade. This refinement can be made more precise as follows (see Sargent and Smith 1997, for a formalization of these ideas).

Suppose that the objects used as money have alternative uses: for example, a coin could be melted down and consumed as metal. Monetary objects may be worth more as money than in their next-best use if their supply is restricted in quantity (by government control over the issue), or unrestricted but at a markup over their alternative value (unlimited minting with a seigniorage charge). Furthermore, these monetary objects, although differing in their alternative values, may have the same value as money if they provide exactly the same monetary services, that is, their relative exchange rate is constant but the level indeterminate. Here, the force that keeps the exchange rate fixed is the Kareken and Wallace (1981) result on indeterminacy. The common value of these monetary objects, as money, will depend on the overall demand for monetary services (the 'needs of trade'), and the total supply of these objects. These conditions may change so as to drive down the value of the objects in monetary use. For example, additions of some monetary objects to the money supply will, if the money demand remains constant, drive down the value of all monetary objects at the same time. If the value falls low enough, the objects with the most valuable alternative use will be the first to be removed from monetary use: the 'best money' will be driven out.

Three points should be noted about this qualified version of Gresham's Law. The foregoing reasoning described one possible equilibrium, but does not rule one where Gresham's Law would fail (such as one money circulating and all the others in their alternative use). Second, the best money can be driven out by the addition of any monetary object to the money supply, not necessarily the worst. Finally, this version depends crucially on the expectations that holders of the monetary objects may have about future rates of return on the different objects: if the value of money is expected to fall further, why would agents persist in holding balances in those objects whose value will fall further than others?

The third refinement arose from the growing importance of asymmetric information in economics. In his famous paper, Akerlof (1970, p. 489) noted the analogy between 'lemons' driving out good cars from the market and Gresham's Law, although he thought that, in the latter case, 'both buyer and seller can tell the difference between good and bad money.' Velde et al. (1999) and Dutu et al. (2005) have explored the role of asymmetric information about coinage. Gresham's Law can make an appearance in such models because of a 'lemons effect'. Good and bad coins will fetch the same price in trades when the seller cannot recognize or be convinced of the quality of the coin being traded, and this may lead to fewer or no trades involving good coins. When both parties do know the difference, they will trade both good and bad coins, with the latter at a premium. This argument, however, rationalizes Gresham's Law only when good and bad coins are

hard to distinguish, for example during medieval debasements in which debased coins were almost identical to the original ones (and thus hard to distinguish) but contained less silver. In most of the cases for which Gresham's Law has been cited, such as milled and hammered money, gold and silver, metal and paper, the monies were clearly distinguishable, and this refinement of Gresham's Law would not apply.

It should now be apparent that trying to salvage Gresham's Law in its lapidary and universal form is hopeless. The phrase would be more useful, not as a general law that excuses the user from providing an explanation, but as a class of outcomes that may or many not obtain, depending on the model's assumptions or the historical episode's circumstances.

Gresham's Law and Bimetallism

Since Gresham's Law owes its fame to its role as an argument against bimetallism during the extensive debates of the late 19th century, a word on this topic is in order. Bimetallism – that is, the concurrent circulation of gold and silver currency at a constant exchange rate – would always be defeated by Gresham's Law: whichever metal was cheaper on the market than at the legal rate would always displace the other. Its application in this context is misplaced. As has been shown (Walras 1977, p. 339; Velde and Weber 2000), bimetallism is a stable monetary system in which exogenous fluctuations (in supply, monetary or non-monetary demand) are accommodated by fluctuations in the relative shares of the two metals in circulation. An increase in the supply of gold need not result in any change in the relative price of the metals, as long as enough gold can be taken out of, and enough silver added to, non-monetary uses, keeping the ratio of marginal utilities constant. This process of displacement of one money by the other takes place precisely as long as the two monies are substitutes at a constant relative price, meaning that neither one can be said to be 'worse' than the other.

See Also

- ▶ Bimetallism
- ▶ Commodity Money
- ▶ Money

Bibliography

Akerlof, G.A. 1970. The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84: 488–500.

Bridrey, É. 1906. *La Théorie de la monnaie au XIVe siècle. Nicole Oresme; étude d'histoire des doctrines et des faits économiques*. Paris: V. Giard & E. Brière.

Burgon, J.W. 1839. *The life and times of Sir Thomas Gresham, Knt., Founder of the Royal Exchange*. Vol. 2. London: Effinghman Wilson.

De Roover, R.A. 1949. *Gresham on foreign exchange: An essay on early English mercantilism with the text of Sir Thomas Gresham's memorandum: For the understanding of the exchange*. Cambridge, MA: Harvard University Press.

Dutu, R., Nosal, E., Rocheteau, G. 2005. On the recognizability of money. Working paper No. 512, Federal Reserve Bank of Cleveland.

Fetter, F.W. 1932. Some neglected aspects of Gresham's law. *Quarterly Journal of Economics* 46: 480–495.

Friedman, M., and A.J. Schwartz. 1963. *A monetary history of the United States, 1867–1960*. Princeton: Princeton University Press.

Giffen, R. 1891. The Gresham law. *Economic Journal* 1: 304–306.

Jevons, W.S. 1875. *Money and the mechanism of exchange*. London: H.S. King & Co..

Kareken, J.H., and N. Wallace. 1981. On the indeterminacy of equilibrium exchange rates. *Quarterly Journal of Economics* 96: 207–222.

Macaulay, T.B., Baron. 1850–1861. The history of England from the accession of James II, 5 London: Longman, Brown, Green and Longmans.

Macleod, H.D. 1855–1856. *The theory and practice of banking: With the elementary principles of currency, prices, credit, and exchanges,* 2 vols. London: Longman, Brown, Green, and Longmans.

Macleod, H.D. 1858. *The elements of political economy*. London: Longman, Brown, Green, Longmans, and Roberts.

Macleod, H.D. 1896. *The history of economics*. New York: Putnam.

Nussbaum, A. 1950. *Money in the law: National and international*. Brooklyn: Foundation Press, Inc.

Palgrave, R.H.I. ed. 1894–1899. *Dictionary of political economy,* 3 vols. London/New York: Macmillan and Co.

Rolnick, A.J., and W.E. Weber. 1986. Gresham's law or Gresham's fallacy? *Journal of Political Economy* 94: 185–199.
Sargent, T.J., and B.D. Smith. 1997. Coinage, debasements, and Gresham's laws. *Economic Theory* 10: 197–226.
Sargent, T.J., and F.R. Velde. 2002. *The big problem of small change*. Princeton: Princeton University Press.
Selgin, G. 2003. Gresham's law. *EH. Net Encyclopedia,* ed. R. Whaples. 10 June. Online. Available at http://eh.net/encyclopedia/article/selgin.gresham.law. Accessed 24 Nov 2006.
Velde, F.R., and W.E. Weber. 2000. A model of bimetallism. *Journal of Political Economy* 108: 1210–1234.
Velde, F.R., W.E. Weber, and R. Wright. 1999. A model of commodity money, with applications to Gresham's law and the debasement puzzle. *Review of Economic Dynamics* 2: 291–323.
Walker, F.A. 1877. *Money.* New York: Henry Holt and Company.
Walras, L. 1977. *Elements of pure economics*. Fairfield: Augustus M. Kelley.
Wolowski, L. 1864. *Traictie de la première invention des monnoies de Nicole Oresme, textes français et latin d'après les manuscrits de la Bibliothèque impériale et Traité de la monnoie de Copernic, texte latin et traduction française*. Paris: Guillaumin.

Griliches, Zvi (1930–1999)

Ernst R. Berndt

Abstract

Born in Lithuania and a survivor of the Dachau concentration camp, Zvi Griliches was one of the most important and influential empirical economists of the second half of the 20th century. Griliches' lifelong research focus involved detailed analyses on the role of technological change as a principal driver of productivity and long-run economic growth. His research contributions were wide-ranging and seminal, and through his students, collaborators and colleagues he greatly affected the conduct of empirical research in economics.

Keywords

American Economic Association; Capital measurement; Diffusion of technology; Discrete count data; Distributed lags; Education and economic growth; Griliches, Z.; Hedonic prices; Inflation; Input–output analysis; Measurement error models; National Bureau of Economic Research; Patents; Pharmaceutical industry; Price measurement; Production functions; Productivity growth; Quality–price relationship; Research and development; Technical change; Unobservable variables

JEL Classifications

B31

Zvi Griliches was one of the most important and influential empirical economists of the second half of the 20th century. His research contributions were wide-ranging and seminal, and through his students, collaborators and colleagues he greatly affected the conduct of empirical research in economics.

Zvi Griliches was born in Lithuania to a well-educated Jewish family. At age 11, along with his parents and sister, Griliches was moved into the ghetto in German-occupied Kaunas, where he remained until 1944, when he and his father were separated from his mother and sister and were sent to Dachau concentration camp. His father died there from starvation, and his mother died from typhus in the Stutthof concentration camp. In May 1945 Griliches was liberated by General Patton's 3rd US Army. After spending a year in Germany, he attempted to go to Palestine, but was prevented by the British from entering, and was held for nine months in an internment camp on Cyprus. He arrived in Palestine in September 1947.

Griliches then spent three years in several kibbutzim, focusing much of his spare time on learning English and mathematics. Although he never went to high school, he taught himself sufficiently well to pass an external high-school equivalence examination in 1950. After studying history for a

year at the Hebrew University, Griliches applied for and was awarded a scholarship by the University of California at Berkeley, where he chose agricultural economics as his major field of study. Griliches completed Berkeley's undergraduate degree requirements in two years, took his first econometrics course from George Kuznets (Simon's brother), and earned his Master's degree in 1954, at which time he transferred to the University of Chicago, where he completed his Ph.D. in 1957, writing a seminal dissertation on the economics of the diffusion of hybrid corn. After 15 years at Chicago, in 1969 Griliches moved to Harvard. In the late 1970s, he became the first Director of the National Bureau of Economic Research's Program on Technological Progress and Productivity Measurement. Elected to the National Academy of Sciences in 1975 and President of the American Economic Association in 1993, Griliches also served as co-editor of *Econometrica* for a decade, and won the John Bates Clark Medal at age 35. He died from cancer in 1999. Further details on his life are found in Lerner (2004) and Trajtenberg and Berndt (2001).

Griliches' lifelong research focus involved detailed analyses on the role of technological change as a principal driver of long-run economic growth, and included examination of the determinants of the diffusion of new technologies, the measurement of physical, human and R&D capital, the role of education, and the contribution of R&D to productivity growth. Griliches devoted a great deal of attention throughout his career to properly measuring various inputs and outputs, and adjusting prices for quality change. Although economic growth theory in the late 1950s emphasized the role of disembodied technological progress and 'manna from heaven', Griliches (1957, 1958, 1960a) instead developed the view that technological progress is itself an economic phenomenon amenable to economic analysis.

In formulating this theme and building supportive empirical evidence, Griliches and his collaborators enlarged considerably the set of econometric tools and procedures now commonly employed by empirical economic researchers. These econometric tool innovations included the use of distributed lags (1961b, 1967a, 1984a), procedures for dealing with measurement error and unobservable variables (1974, 1975, 1977, 1978), and with discrete count data (1984b).

Much of Griliches' early empirical research focused on measuring inputs, outputs and productivity in agriculture (1960a, 1964), but later on it extended to other sectors of the economy, particularly the service sectors (1992, 1994a, 2000a). The theoretical framework integrating these measurement issues involved use of the production function (1964, 1967b, 1969, 1970, 1971). He initially focused on measurement of physical capital inputs (1963, 1966, 1984a), but subsequently on issues involving measurement of labour input that generated several influential strands of research. One important literature involved establishing relationships between quality-adjusted labour inputs, education, and rates of return to schooling (1975, 1979, 1986a, 1997, 2000b). A related literature examined complementarity between physical capital and skilled labour (1969, 1977, 1994b).

In addition to examining the relationships among outputs and labour and capital inputs, a great deal of Griliches' research focused on the special role of R&D, differences between private and social returns to R&D, and the impact of R&D on productivity growth (1958, 1964, 1980, 1994, 1998). This research then led to a more detailed analysis of R&D, including a host of important studies that examined the extent to which patents served as a useful indicator of inventive activity generated by R&D (1986b, c, 1987, 1990).

Another of Griliches' seminal contributions involved reviving intellectual and policy interest in the use of hedonic multivariate regression techniques to adjust observed prices for changes in observed and unobserved quality over time. In large part this research reflected Griliches' scepticism about measuring productivity growth as the 'residual', which in its simplest form simply subtracted the growth of traditionally measured inputs from the growth of outputs. To what extent, he reasoned, could what goes into that residual ('a measure of our ignorance') reflect instead measurement and specification errors, rather than technological and other quality change? This led Griliches to undertake empirical analyses initially

linking prices of new automobiles to observed characteristics (1961a), as well as prices of used automobiles to operating cost characteristics, such as fuel efficiency and gasoline prices (1986d). Subsequent research examined the extent to which traditionally constructed government price statistics for certain high-technology goods such as personal computers overstated price inflation (or understated price deflation) by failing to incorporate fully the quality attribute improvements that were embodied in new goods (1993a, 1995). Yet another strand of this hedonic price research extended into prescription pharmaceuticals (1996), where problems of a new goods bias and oversampling of older goods were particularly important (1993b, 1994c). Adjusting measures of medical price inflation for changes in outcomes from medical treatments was a focus of Griliches' work with collaborators just prior to his death in 1999 (2000a).

Griliches' interest in price measurement led to his being appointed a member of both the 1960–1961 Stigler Commission (resulting in 1961a), and the Boskin Commission of 1995–1996, each of which provided influential and thoughtful recommendations on price measurement issues facing the U.S. Bureau of Labor Statistics. Although he was also named a member of the National Academy of Science Panel on the Conceptual, Measurement and Other Statistical Issues in Developing Cost-of-Living Indexes (1999–2001), because of his ill health and subsequent death in late 1999 he was unable to contribute directly to the final report.

See Also

- ▶ Agricultural Economics
- ▶ Distributed Lags
- ▶ Econometrics
- ▶ Ghettoes
- ▶ Growth Accounting
- ▶ Hedonic Prices
- ▶ Level Accounting
- ▶ Measurement Error Models
- ▶ Production Functions
- ▶ Technology

Selected Works

1957. Hybrid corn: An exploration in the economics of technological change. *Econometrica* 25: 501–522.

1958. Research cost and social returns: Hybrid corn and related innovations. *Journal of Political Economy* 66: 419–431.

1960a. Measuring inputs in agriculture: A critical survey. *Journal of Farm Economics* 42: 1411–1433.

1960b. Hybrid corn and the economics of innovation. *Science* 132: 275–280.

1961a. Hedonic price indexes for automobiles: An econometric analysis of quality change. In *The price statistics of the federal government.* General Series No. 73. New York: NBER.

1961b. A note on serial correlation bias in estimates of distributed lags. *Econometrica* 29: 65–73.

1963. Capital stock in investment functions: Some problems of concept and measurement. In *Measurement in economics: Studies in mathematical economics and econometrics in memory of Yehuda Grunfeld,* ed. C. Christ et al. Stanford: Stanford University Press.

1964. Research expenditures, education, and the aggregate agricultural production function. *American Economic Review* 54: 961–974.

1966. (With D. Jorgenson.) Sources of measured productivity change: Capital input. *American Economic Review* 56(2): 50–61.

1967a. Distributed lags: A survey. *Econometrica* 35: 16–49.

1967b. (With D. Jorgenson.) The explanation of productivity change. *Review of Economic Studies* 34: 249–283.

1969. Capital skill complementarity. *Review of Economics and Statistics* 51: 465–468.

1970. Notes on the role of education in production functions and growth accounting. In *Education, income, and human capital*, ed. W. Hansen. NBER Studies in Income and Wealth, vol. 35. New York: Columbia University Press.

1971. (With V. Ringstad.) *Economies of scale and the form of the production function*. Amsterdam: North-Holland.

1974. Errors in variables and other unobservables. *Econometrica* 42: 971–978.

1975. (With G. Chamberlain.) Unobservables with a variance components structure: Ability, schooling, and the economic success of brothers. *International Economic Review* 16: 422–429.

1977. Estimating the returns to schooling: Some econometric problems. *Econometrica* 45: 1–22.

1978. (With B. Hall, and J. Hausman.) Missing data and self-selection in large panels. *Annales de l'INSEE* 30–31: 137–176.

1979. Sibling models and data in economics: Beginnings of a survey. *Journal of Political Economy* 87(5, Part II): S37–S64.

1980. R&D and the productivity slowdown. *American Economic Review Papers and Proceedings* 70: 343–348.

1984a. (With A. Pakes.) Estimating distributed lags in short panels with an application to the specification of depreciation patterns and capital stock constructs. *Review of Economic Studies* 51: 243–262.

1984b. (With J. Hausman, and B. Hall.) Econometric models for count data with an application to the patent R&D relationship. *Econometrica* 52: 909–938.

1986a. (With J. Bound, and B. Hall.) Wages, schooling and IQ of brothers and sisters: Do the family factors differ? *International Economic Review* 27: 77–105.

1986b. Economic data issues. In *Handbook of econometrics,* ed. Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

1986c. (With B. Hall, and J. Hausman.) Patents and R&D: Is there a lag? *International Economic Review* 27: 265–283.

1986d. (With M. Ohta.) Automobile prices and quality change: Did the gasoline price increases change consumer tastes in the U.S.? *Journal of Business and Economic Statistics* 4: 187–198.

1987. (With A. Pakes, and B. Hall.) The value of patents as indicators of inventive activity. In *Economic policy and technological performance*, ed. P. Dasgupta and P. Stoneham. Cambridge: Cambridge University Press.

1990. Patent statistics as economic indicators: A survey. *Journal of Economic Literature* 28: 1661–1707.

1992. (Ed.) *Output measurement in the service sectors.* NBER Studies in Income and Wealth, vol. 56. Chicago: University of Chicago Press.

1993a. (With E. Berndt.) Price indexes for microcomputers: An exploratory study. In *Price measurements and their uses*, ed. M. Foss, M. Manser, and A. Young. NBER Studies in Income and Wealth, vol. 57. Chicago: University of Chicago Press.

1993b. (With E. Berndt, and J. Rosett.) Auditing the producer price index: Micro evidence from prescription pharmaceutical preparations. *Journal of Business and Economic Statistics* 27: 131–152.

1994a. Productivity, R&D and the data constraint. *American Economic Review* 84: 1–23.

1994b. (With E. Berman, and J. Bound.) Changes in the demand for skilled labor within U.-S. manufacturing industries: Evidence from the Annual Survey of Manufacturing. *Quarterly Journal of Economics* 109: 367–398.

1994c. (With I. Cockburn.) Generics and new goods in pharmaceutical price indexes. *American Economic Review* 84: 1213–1232.

1995. (With E. Berndt, and N. Rappaport) Econometric estimates of price indexes for personal computers in the 1990s. *Journal of Econometrics* 68: 243–268.

1996. (With E. Berndt, and I. Cockburn) Pharmaceutical innovations and market dynamics: Tracking effects on price indexes for antidepressant drugs. *Brookings Papers on Economic Activity: Microeconomics* 1996(2): 133–188.

1997. Education, human capital and growth: A personal perspective. *Journal of Labor Economics* 15(1, Part II): S330–S344.

1998. *R&D and productivity: The econometric evidence.* Chicago: University of Chicago Press.

2000a. (With E. Berndt, D. Cutler, R. Frank, J. Newhouse, and J. Triplett.) Medical care prices and output. In *Handbook of health economics,* vol. 1A, ed. J. Newhouse and A. Culyer. Amsterdam: Elsevier Science.

2000b. *R&D, education and productivity: A retrospective.* Cambridge, MA: Harvard University Press.

Bibliography

Lerner, B. 2004. From the particular to the aggregate: The story of Zvi Griliches. In *The triumph of wounded souls: Seven Holocaust survivors' lives.* Notre Dame: University of Notre Dame Press.

Trajtenberg, M., and E. Berndt. 2001. In Memoriam: Zvi Griliches (1930–1999). *Journal of Economic and Social Measurement* 27: 93–97.

Gross Substitutes

Lionel W. McKenzie

Abstract

The gross substitute assumption is used to establish the existence and uniqueness of an equilibrium and to prove the equilibrium to be stable for a dynamic adjustment system for prices. The gross substitute assumption also implies results of comparative statics, that is, results on the displacement of equilibrium that follows from shifts in demand or changes in initial stocks. The concept of gross substitutes was introduced by Mosak (*General equilibrium theory in international trade*. Bloomington: Principia Press, 1944) in the context of a pure trading model. A definition with wider application is the one used by Morishima (*Equilibrium, stability and growth*. Oxford: Clarendon Press, 1964).

Keywords

Comparative statics; Excess demand; Existence of equilibrium; Global stability theorems; Gross substitutes; Income effect; Revealed preference; Stability; Substitution effect; Tâtonnement; Temporary equilibrium; Uniqueness of equilibrium; Walras's Law; Weak gross substitutes

JEL Classifications

D0

The assumption that goods are gross substitutes is applied to a set of excess demand functions $e_i(p_1, \ldots, p_n)$, $i = 1, \ldots, n$, where p_i is the price of the ith good and e_i is the excess demand for the ith good. The concept was introduced by Mosak (1944) in the context of a pure trading model. However, his definition required that $\partial e_i|\partial p_j$ have the same sign as the substitution term in the Slutsky equation as well as be of positive sign. That is, the income effect should not overbalance the substitution effect. At about the same time Metzler (1945) said simply that the jth good is a gross substitute for the ith good if $e_{ij} = \partial e_i(p)|\partial p_j > 0$ holds, and this has been the meaning used in later papers when the functions e_i have been assumed to be differentiable.

A definition with wider application is the one used by Morishima (1964). By this definition the jth good is a gross substitute for the ith good if $e_i(p) < e_i(p')$ whenever $p \leq p', p_j < p'_j, p'_k = p_k$ for $k \neq j$. We will say that the assumption of gross substitutes holds if $e_i(p) < e_i(p')$, $i = 1, \ldots, n$, for all p and p' such that $p \leq p'$, $p \neq p'$, and $p_i = p'_i$. If $e(p)$ is *differentiable*, this assumption implies that the $n \times n$ matrix $[e_{ij}(p)]$ has positive off-diagonal elements for all p in the interior of the domain of e (p). A matrix with this property and a negative diagonal is often referred to as a Metzler matrix. We will say that the assumption of weak gross substitutes holds if the condition $e_j(p) < e_i(p')$ is replaced by the weak inequality $e_i(p) \leq e_i(p')$. In the case of differentiable $e(p)$ it is implied by the assumption of weak gross substitutes that $[e_{ij}(p)]$ has non-negative off-diagonal elements. That these assumptions are not empty is shown by the case of excess demands defined by Cobb–Douglas utility functions of the form $U(x) = \prod_{i=1}^{n} x_i^{\alpha_j}$, where $\alpha_i > 0$ and $\sum \alpha_j = 1$. The excess demand function for a consumer in a pure exchange economy, holding initial stocks $\bar{x}$, is $e_j(p) = \left(\alpha_i \sum_{k=1}^{n} p_k \bar{x}_k / p_i\right) - \bar{x}_i$, so $e_{ij}(p) = \alpha_i \bar{x}_j / p_j \geq 0$ for $i \neq j$. If the initial stock of every good is positive for the whole market, the assumption of gross substitutes is satisfied.

A price vector p is said to be an equilibrium of the set of demand functions if $e(p) \leq 0$. The gross substitute assumption is used to establish the existence and uniqueness of an equilibrium and to prove the equilibrium to be stable for a dynamic adjustment system for prices. The gross substitute assumption also implies results of comparative statics, that is, results on the displacement of equilibrium that follows from shifts in demand or changes in initial stocks.

The following assumptions will be made on (p).

(B) $e(p)$ is defined for all $p > 0$, and $p^s \to p$ where $p_i = 0$ for $i \in I$, $p \neq 0$ implies $\sum_{i \in I} |e_i(p)| \to \infty$. Also $e(p)$ is bounded below.

(C) $e(p)$ is single-valued and continuous for $p > 0$.

(H) $e(p) = e(\lambda p)$ for any $\lambda > 0$, that is $e(p)$ is positively homogeneous of degree 0.

(W) $\sum_{i=1}^{n} p_i e_i(p) = 0$, that is, $e(p)$ satisfies Walras's Law.

The example of Arrow-Hahn (1971, pp. 29–30), where demand is derived from the utility function $u(x) = x_1^{1/2} + x_2^{1/2} + x_3^{1/2}$ where x is the consumption bundle, shows that assumption B cannot easily be improved.

Uniqueness

The existence of a positive equilibrium under assumptions B, C, H, and W does not require an assumption of gross substitutes (see Debreu 1970). However, if thereshould exist an equilibrium price p when gross substitutes is assumed, $p > 0$ musthold. This follows from the fact that $p_i = 0$ for some i, and $p \neq 0$, is inconsistent with assumption H in that case, since it must at the same time be true, for $\lambda > 1$, that $e_i(\lambda p) = e_i(p)$ and $e_i(\lambda p) > e_i(p)$. Thus $e(p)$ cannot be defined at such a p.

Make the assumption that all goods are gross substitutes. Assume that two equilibria exist, say p and p' where $p' \neq \lambda p$ for any $\lambda > 0$. By assumption H we may chose p and p' so that $p_i = p'_i$ for some i and $p_j \leq p'_j$ for $j \neq i$. Then by gross substitutes $e_i(p) < e_i(p')$. But p and $p' > 0$ and assumption W implies that $e_i(p) = e_i(p') = 0$, which is a contradiction. Thus an equilibrium price vector is unique up to multiplication by a positive number.

Consider a partition of the goods into two non-empty subsets I and J. Say that the excess demand functions are connected if for any such partition $p_i = p'_i$ for all $i \in I$ and $p_j < p'_j$ for all $j \in J$ implies that $(p) \neq e(p')$. Then by a similar argument, uniqueness of equilibrium may be seen to hold when weak gross substitutes and connectedness are assumed. Strictly speaking, connectedness need only hold at an equilibrium point.

If weak gross substitutes is assumed without connectedness, uniqueness may fail. However, it may be shown that the set of equilibrium price vectors is convex (McKenzie 1960). Arrow and Hurwicz (1960) proved that the weak axiom of revealed preference holds between any equilibrium price vector p and any non-equilibrium price vector p' when weak gross substitutes is assumed. This means that $0 = p'e(p) = p'e(p')$ implies $pe(p') > pe(p) = 0$. In other words, $pe(p') > 0$. Suppose p and p' are both equilibria and consider $p'' = \alpha p + (1 - \alpha)p'$ for $0 < \alpha < 1$. If p'' is not an equilibrium $p''e(p'') = \alpha pe(p'') + (1 - \alpha)p'e(p'') > 0$ which conradicts assumption W, Walras's Law. Thus p'' is an equilibrium and the set of equilibria is convex.

Comparative Statics

The modern approach to the comparison of equilibria after a shift of demand was begun by Hicks (1939). The fact that the Hicksian theorems hold locally when the excess demand functions satisfy the gross substitute assumption was proved by Mosak (1944). A global treatment of comparative statics in this context was given by Morishima (1964). Assume weak gross substitutes. Let demand shift from the *ith* good to the *j*th good. Let p and p' be the old and new equilibrium prices, e and e' the old and new excess demand functions. Then

$$p_i'e_i'(p) + p_j'e_j'(p) = \sum_{i=1}^{n} p_k'e_k'(p) > 0, \text{by the Weak Axiom.} \quad (1)$$

On the other hand,

$$p_ie_i'(p) + p_je_j'(p) = \sum_{i=1}^{n} p_ke_k'(p) = 0, \text{by Walras}' \text{ Law.} \quad (2)$$

Multiply (1) by p_i and (2) by p_i' and subtract to obtain $p_ip_j' - p_i'p_j > 0$, or $p_j'/p_j > p_i'/p_i$. Thus the price of the *j*th good increases relative to the price of the *ith* good.

Assume that all goods are gross substitutes. Then a shift in demand from the *ith* good to the *j*th good raises the price of the *i*th good relative to all other goods and lowers the price of the *ith* good relative to all other goods. These results areimmediate from the fact that the good that rises in price relative to some goods and falls relative to none must experience a fall in demand and the good that falls in price relative to some goods and rises relative to none must experience a rise in demand. But only the *j*th good can absorb a rise and only the *i*th good can absorb a fall and still have demand equal to 0 after the shift has occurred. All other goodshave zero excess demands at the old equilibrium prices after the shift, and the excess demands at the new equilibrium prices must also be zero. The same results follow from weak gross substitutes if any subset of the excess demand functions with $n - 1$ members is connected.

If the $e_i(p)$ are assumed to be continuously differentiable, the local theory of comparative statics is equivalent to determining the sign pattern of the inverses of the principal submatrices of order $n - 1$ of the Jacobian $[e_{ij}]i, j = 1, \ldots, n$. The gross substitutes assumption implies that $e_{ij} > 0$ for $i \neq j$. Then either assumption H or W implies that the inverses of these submatrices have all elements negative. Choose the *n*th good as numeraire, and choose units so equilibrium prices $p_j = 1$, all *i*. Then $dp_j/d\alpha = -\left([e_{ij}]_{nn}\right)^{-1}_{ih}$ if $\partial e_h(p, \alpha)/\partial\alpha = 1$. This is minus the *ith* element of the *h*th column of the inverse matrix of the submatrix where the *nth* row and column are omitted. Thus $dp_i/d\alpha > 0$, and it may be shown that $dp_h/d\alpha > dp_j/d\alpha$ for $i \neq h$ or *n*.

If weak gross substitutes is assumed but the Jacobian and its $n - 1$ principal minors are indecomposable the same conclusions follow.

The local results may be extended to the case where there is a numeraire and the goods other than the numeraire may be partitioned into two non-empty subsets with indices in *I* and *J*, such that $e_{ij} > 0$ for $i \neq j$ and *i* and *j* in the same subset, while $e_{ij} < 0$ for *i* and *j* in different subsets. If the principal minors of $[e_{ij}]$ of order $n - 1$ have dominant diagonals, at equilibrium with the equilibrium prices as multipliers, the shifts of demand raise the price of the good to which demand has shifted and also the prices of all other goods in the subset of the partition to which it belongs, while lowering the prices of the goods in the other subset. Also the beneficiary of the demand shift has the largest change in equilibrium price in absolute value. These results are seen to follow from those for gross substitutes by considering the matrix formed by pre- and post-multiplying a principal minor of $[e_{ij}]$ by a diagonal matrix *D* with $d_{ii} = 1$ for $i \in I$ and $d_{jj} = -1$ for $j \in J$. This case was first analysed byMorishima (1952). The gross substitute case and the Morishima case may be shown to be the only sign patterns for a Jacobian matrix of the demand functions with all elements non-zero which allow the inverse matrix to be signed without quantitative information (see Bassett et al. 1967).

Stability

Since weak gross substitutes implies the weak axiom of revealed preference which in turn implies local stability of the tâtonnement for both the usual price adjustment models, with and without a numeraire (Arrow and Hurwicz 1958), there is no special advantage for gross substitutes in local analysis of stability. However, for global results on stability the gross substitute assumptions are the only ones known with much

plausibility. In order to use the weak axiom the adjustment must be, after proper choice of units,

$$\dot{p}_i = e_j(p), \tag{I}$$

for all goods other than the numeraire, if any, where $\dot{p}_i = \partial p_i / \partial t$, and (I) must hold globally. Excess demand is now assumed to be continuously differentiable. In order to use the other major possibility, a dominant diagonal for the matrix of demand elasticities, assuming a numeraire, the adjustment process

$$\dot{p}_i = p_i e_i(p), \tag{II}$$

is used for the non-numeraire goods (see Arrow and Hahn 1971, p. 293). While (II) may be more reasonable than (I) for a global adjustment rule it is also very special.

On the other hand, when weak gross substitutes is assumed, stability may be proved for the adjustment rule

$$\dot{p} = h_i(p), \tag{III}$$

for all goods other than the numeraire, if any. The only special requirements placed on $h_i(p)$ are that it should be continuously differentiable and have the sign of $e_i(p)$. The adjustment rule III was proposed by McKenzie (1960) and global stability was proved using as a Lyapunov function the value of positive excess demand, $V[p(t)] = \sum_{i \in P} p_i e_i(p)$, $P = \{i | \ e_i(p) \geq 0\}$. The tâtonnement is shown to converge to the convex set of equilibrium prices. If the excess demand functions are connected, the equilibrium is unique.

Arrow and Hurwicz (1962) proved that the convergence of process III with weak gross substitutes is, in fact, to a particular equilibrium price vector, which will depend on the initial prices. This is clear once it is recognized that the goods whose prices are highest relative to some equilibrium price cannot be in excess demand and their prices cannot rise, and *mutatis mutandis* for the prices which are lowest. Thus the prices during tâtonnement cannot retreat from any equilibrium price vector. In the case of gross substitutes this line of argument is very simple and effective, since the prices actually fall and rise respectively.

In the theory of tâtonnement prices are revised according to excess demand but no trading occurs until equilibrium is reached. However, the stability of the adjustment process for a pure exchange economy is not lost if trading occurs at market prices, so long as the excess demand that drives the tâtonnement is determined by the maximization of utility by each trader under a budget equal to the value of the stocks he currently holds (Negishi 1961). The crucial fact is that trading at market prices has only second-order effects on excess demand. However, the price to which the process converges now depends on initial conditions and the course of trading.

It was pointed out by Rader (1972) that the production sector of the economy is unlikely to satisfy the gross substitute assumption in the demand for factors. As a consequence it seemed that the range of application of the gross substitutes assumption was effectively confined to pure trading economies. However, Rader was able to prove a local stability theorem assuming gross substitutes only for households. The production sector is made up of a finite number of firms with strictly convex production possibility sets. The key to the argument is that $p\left[e_{ij}^{H}\right] = 0$ at equilibrium, where $e^{\mathrm{H}}(p)$ is household excess demand. This is established by differentiating Walras's Law $pe^{F}(p) + pe^{H}(p) = 0$ to give

$$e^{\mathrm{F}}(p) + e^{\mathrm{H}}(p) + p\left[e_{ij}^{\mathrm{F}} + pe_{ij}^{\mathrm{H}}\right] = 0, \tag{3}$$

Since the first two terms sum to 0 at equilibrium, and the third term is 0 by profit maximization, (3) implies $p\left[e_{ij}^{\mathrm{H}}\right] = 0$. Then the Jacobian of the adjustment system I with a numeraire is negative definite at equilibrium and local stability follows.

Generalizations

Mukherji (1972) pointed out that some gross substitute theorems carry over if the weak gross

substitute pattern is established in a transformed goods space. In particular, if there exists a matrix S such that $S^{-1}[e_{ij} + e_{ji}]S$ is indecomposable with off-diagonal elements non-negative, the tâtonnement is locally stable for the process I, with or without a numeraire. Also a rise in demand for the *ith* good causes the *ith* equilibrium price to rise. Ohyama (1972) shows that similar results follow if there is a stochastic matrix G which is positive definite and $G[e_{ij}]$ or $G[e_{ij} + e_{ji}]$ satisfies the conditions above. Uzawa (1960) proved that a discrete tâtonnement defined by $P_i(t+1) = \max\{0, p_i(t) + f_i(t)\}$ where $f_i(t) = \beta_i\{e_i[p(t)]\}$ is globally stable when the Weak Axiom holds and $\beta_i > 0$ is sufficiently small. He assumes a numeraire and some additional differentiality and nonsingularity conditions.

Howitt (1980) defines a generalized gross substitute notion where $e(p)$ is allowed to be a convex valued correspondence rather than a function. This assumption of generalized gross substitutes holds if

(a) for all p, $p' > 0$, if there is a partition of indices for goods into non-empty subsets I and J where $p'_i = p_i$ for all $i \in I$ and $p'_j > p_j$ for all $j \in J$ than $\sum_{i\in I} p_i x'_i \geq \sum_{i\in I} p_i x_i$ for all $x \in e(p)$, $x' \in e(p')$;
(b) strict inequality holds in (a) if p is an equilibrium.

Howitt proves that the equilibrium price vector is unique and globally stable for the price adjustment process

$$\dot{p}_i \in e(p), \tag{IV}$$

under generalized gross substitutes, when assumptions C , W, and H hold and an equilibrium exists. Excess demand $e(p)$ is assumed to be upper semicontinuous. He applies his result to the linear economy described by Gale (1976) and shown to satisfy gross substitutes by Cheng (1979).

Arrow and Hurwicz (1962) extended the adjustment process III to include expected prices. Let q represent expected prices. Then their process with adaptive expectations is

$$\dot{p}_i = h_i(p/q), \text{ if } p_i > 0 \text{ and } h_i(p, q) > 0, \ = 0, \text{ otherwise}$$

$$\begin{aligned} &\text{sing } h_i(p,q) = \text{sing}\, e_i(p,q) \text{ or } = 0 \\ &\text{if } i \text{ is a numeraire } \dot{q}_i = a_i(q_i - p_i) \end{aligned} \tag{V}$$

They prove global stability for this process under weak gross substitutes with some auxiliary assumptions. Arrow and Hahn (1971) give a proof of global stability with an assumption of gross substitutes and assumptions B, C, H, and W. By the gross substitute assumption in this context is meant that $\partial e_i(p, q)/\partial p_j > 0$ for $i \neq j$, and $\partial e_i(p, q)/\partial q_j > 0$ for all j. The adjustment function h_i is assumed to be continuously differentiable.

Arrow and Hahn also prove a global stability theorem for a model in which expected prices q_i are given as functions $q_i(p)$ of current prices. This is in accord with models of temporary equilibrium. They prove global stability for adjustment process II with a numeraire, assuming gross substitutes for $e(p, q)$ and the Hicksian elasticity of substitution $\varepsilon_j = d \log q_j / d \log p_j \leq 1$, which is consistent with Hicks's presumption when the strict inequality holds (see Hicks 1939, p. 251).

In this model of the tâtonnement for temporary equilibrium $q(p)$ is presumably the expected price on the assumption that p is an equilibrium price. It is not so clear how to justify adaptive expectations in the tâtonnement setting.

See Also

▶ Substitutes and Complements

Bibliography

Arrow, K.J., and F.H. Hahn. 1971. *General competitive analysis*. San Francisco: Holden-Day.
Arrow, K.J., and L. Hurwicz. 1958. On the stability of the competitive equilibrium, I. *Econometrica* 26: 522–552.
Arrow, K.J., and L. Hurwicz. 1960. Competitive stability under weak gross substitutability: The 'Euclidean distance approach'. *International Economic Review* 1: 38–49.
Arrow, K.J., and L. Hurwicz. 1962. Competitive stability under weak gross substitutability: Nonlinear price

adjustment and adaptive expectations. *International Economic Review* 3: 233–255.
Bassett, L., H. Habibagahi, and J. Quirk. 1967. Qualitative economics and Morishima matrices. *Econometrica* 35: 221–233.
Cheng, H.-C. 1979. Linear economies are 'gross substitute' systems. *Journal of Economic Theory* 20: 110–117.
Debreu, G. 1970. Economies with a finite set of equilibria. *Econometrica* 38: 387–392.
Gale, D. 1976. The linear exchange model. *Journal of Mathematical Economics* 3: 205–259.
Hicks, J.R. 1939. *Value and capital*. Oxford: Clarendon Press.
Howitt, P. 1980. Gross substitutability with multi-valued excess demand functions. *Econometrica* 48: 1567–1575.
McKenzie, L.W. 1960. Stability of equilibrium and the value of positive excess demand. *Econometrica* 28: 606–617.
Metzler, L. 1945. Stability of multiple markets: The Hicks conditions. *Econometrica* 13: 277–292.
Morishima, M. 1952. On the laws of change of price-system in an economy which contains complementary commodities. *Osaka Economic Papers* 1: 101–113.
Morishima, M. 1964. *Equilibrium, stability and growth*. Oxford: Clarendon Press.
Mosak, J.L. 1944. *General equilibrium theory in international trade*. Bloomington: Principia Press.
Mukherji, A. 1972. On complementarity and stability. *Journal of Economic Theory* 4: 442–457.
Negishi, T. 1961. On the formation of prices. *International Economic Review* 2: 122–126.
Ohyama, M. 1972. On the stability of generalized Metzlerian systems. *Review of Economic Studies* 39: 193–204.
Rader, T. 1972. General equilibrium theory with complementary factors. *Journal of Economic Theory* 4: 372–380.
Uzawa, H. 1960. Walras' tâtonnement in the theory of exchange. *Review of Economic Studies* 27: 182–194.

Grossman, Herschel I. (1939–2004)

Enrico Spolaore

Keywords

Excess demand and supply; Grossman, H.; Haavelmo, T.; Keynesianism; Labour demand; Labour markets; Non-clearing markets; Patinkin, D.

JEL Classifications

B31

Herschel I. Grossman was born in Philadelphia on 6 March 1939. He obtained a BA from the University of Virginia in 1960, a B.Phil. from Oxford in 1962, and a Ph.D. from Johns Hopkins in 1965. He joined Brown University's economics faculty in 1964. He died suddenly while attending a conference in Marseilles on 9 October 2004.

In the early 1970s Grossman, in cooperation with Robert J. Barro, produced path-breaking research on the foundations of Keynesian macroeconomics. Barro and Grossman's main joint contribution was the article 'A General Disequilibrium Model of Income and Employment' (1971a). This was the first formalization linking the labour market and the output market in a general-disequilibrium setting with exogenous wage rate and price level, in which labour demand is constrained by output sales while the demand for goods is constrained by sales in the labour market. Generalized excess supply implied a Keynesian demand-multiplier effect. The analysis also shed light on cases of generalized excess demand, such as in socialist economies with artificially low prices. This work unified complementary strands of the literature – particularly, Patinkin (1965, ch. 13) and Clower (1965) – within a coherent choice-theoretic framework, and for many years was the most cited article ever published in the *American Economic Review.* Barro and Grossman summarized and extended their analysis in the landmark book *Money, Employment, and Inflation* (1976). Independent appraisals of the non-market-clearing paradigm and its limits were given by Barro (1979) and Grossman (1979). See also Grossman's entry on Monetary Disequilibrium and Market Clearing in the first (1987) edition of *The New Palgrave: A Dictionary of Economics.*

In his subsequent work Grossman became increasingly interested in understanding the foundations of economic policy and the effects of conflict on the economy, and made innovative contributions to political economy and the economics of appropriation (see Kolmar 2005, for a survey). Building on Haavelmo (1954), Grossman modelled conflict as an economic activity by agents allocating resources over various uses, including the uses for appropriative conflict itself.

His contributions in this area comprise theories of governments as kleptocracies (1990, 1994a, 1999), insurrections (1991), appropriation and land reform (1994b), effective property rights (1995), anarchy, predation and the state (2002), and many others.

See Also

- ▶ Defence Economics
- ▶ Fixprice Models
- ▶ Haavelmo, Trygve (1911–1999)
- ▶ Keynesianism
- ▶ Patinkin, Don (1922–1955)

Selected Works

1971a. (With R. Barro.) A general disequilibrium model of income and employment. *American Economic Review* 61: 82–93.
1971b. Money, interest, and prices in market disequilibrium. *Journal of Political Economy* 79: 943–961.
1974. (With R. Barro.) Suppressed inflation and supply multiplier. *Review of Economic Studies* 41: 87–104.
1976. (With R. Barro.) *Money, employment, and inflation.* New York: Cambridge University Press.
1978. Risk shifting, layoffs, and seniority. *Journal of Monetary Economics* 4: 661–686.
1979. Why does aggregate employment fluctuate? *American Economic Review* 69: 64–69.
1987. Monetary disequilibrium and market clearing. In *The new Palgrave: A dictionary of economics,* vol. 3, ed. J. Eatwell, M. Milgate, and P. Newman. Basingstoke: Palgrave.
1988a. (With B. Diba.) Explosive rational bubbles in stock prices? *American Economic Review* 78: 520–530.
1988b. (With J. Van Huyck.) Sovereign debt as a contingent claim. excusable default, repudiation, and reputation. *American Economic Review* 78: 1088–1097.
1990. (With S. Noh.) A theory of kleptocracy with probabilistic survival and reputation. *Economics and Politics* 2: 157–171.
1991. A general equilibrium model of insurrections. *American Economic Review* 81, 912–921.
1994a. (With S. Noh.) Proprietary public finance and economic welfare. *Journal of Public Economics* 53: 187–204.
1994b. Production, appropriation, and land reform. *American Economic Review* 84: 705–712.
1995. (With M. Kim.) Swords or plowshares? A theory of the security of claims to property. *Journal of Political Economy* 103: 1275–1288.
1999. Kleptocracies and revolutions. *Oxford Economic Papers* 51: 267–283.
2002. 'Make us a king': Anarchy, predation, and the state. *European Journal of Political Economy* 18: 31–46.

Bibliography

Barro, R. 1979. Second thoughts on Keynesian economics. *American Economic Review* 69: 54–59.
Clower, R. 1965. The Keynesian counterrevolution: A theoretical appraisal. In *The theory of interest rates*, ed. F. Hahn and F. Brechling. London: Macmillan.
Haavelmo, T. 1954. *A study in the theory of economic evolution*. Amsterdam: North-Holland.
Kolmar, M. 2005. The contribution of Herschel I. Grossman to political economy. *European Journal of Political Economy* 21: 802–814.
Patinkin, D. 1965. *Money, interest and prices*, 2nd ed. New York: Harper and Row.

Grossmann, Henryk (1881–1950)

Josef Steindl

Abstract

Grossmann was born on 14 April 1881 in Cracow and died on 24 November 1950 in Leipzig. He studied in Cracow and lived from 1908 to 1918 in Vienna (collaborating with Carl Grünberg) and 1918 to 1925 in Warsaw (at the Central Statistical office and the Free University). From 1925 to 1933 he was a political refugee in Germany (University of

Frankfurt/Main) and later in France, England and USA. Grossmann spent his last years in the German Democratic Republic, at the University of Leipzig.

Grossmann was born on 14 April 1881 in Cracow and died on 24 November 1950 in Leipzig. He studied in Cracow and lived from 1908 to 1918 in Vienna (collaborating with Carl Grünberg) and 1918 to 1925 in Warsaw (at the Central Statistical office and the Free University). From 1925 to 1933 he was a political refugee in Germany (University of Frankfurt/Main) and later in France, England and USA. Grossmann spent his last years in the German Democratic Republic, at the University of Leipzig.

His main work (1929), based on Marx, deals with the inevitability of the breakdown of capitalism. The method of Marx, in his view, is a stepwise approximation to reality which starts from a simplified abstract model of the accumulation process (based on the reproduction schema and assuming a closed system, two classes only, no credit, commodities sold at their values, constant value of money) and proceeds by gradually adding realistic details of secondary importance ('surface phenomena') among which he counts monopoly, money and credit, capital exports and the struggle for raw materials. In following this method Grossmann demonstrates the inevitability of breakdown and then deals with the factors which counteract and therefore delay the breakdown.

The starting point of his theory is an arithmetic example of Otto Bauer based on the reproduction schema of Marx which was intended by Bauer (in a polemic against Rosa Luxemburg) to demonstrate that realization under extended reproduction was perfectly possible. Bauer worked out his example only for four years, but Grossmann extended it to 35 years in order to demonstrate that the accumulation process could not proceed without limit. Following Bauer, he made the following assumptions: 5 per cent growth of variable capital (determined exogenously by the growth of population) while the constant capital was to grow by 10 per cent; surplus value was to be constant at 100 per cent.

Since the organic composition of capital was continuously increasing, the Marxian conclusion of a declining rate of profit held good, as also shown in Bauer's example. This implied, with constant growth of capital, that the share of consumption in surplus value had to decrease. Grossmann sees no difficulty in this so long as the absolute amounts of profit and consumption increase (owing to the increase in capital). For this reason he considers Marx's theory as incomplete. His own contribution is to show that the *absolute* amounts also have to decrease. The step taken from Marx seems to be fairly simple: if the declining rate of profit is combined with an exogenously given constant rate of growth of capital, then the share of consumption in surplus must ultimately go down to zero and below. This marks the point of breakdown.

Grossmann deals extensively with counteracting tendencies such as new colonial markets and real wage cuts. These can only delay and not avoid the breakdown. Their effect will appear in the form of cyclical crises which Grossmann expected would become more and more serious.

Grossmann strongly criticizes all Marxist writers before him (in particular, Hilferding and Luxemburg) for having distorted the content of Marx's teaching. His aim is to restore the orthodoxy of the true Marx which in his view is embodied in the breakdown thesis based on the increase in the organic composition of capital. Other aspects of Marx he plays down (historical materialism) or ignores (the realization problem). Since his book largely takes the form of a polemic it may be used as a source of information on Marxist and other literature, but he does scant justice to the ideas of some of these writers.

Grossmann was, however, a man of culture and learning, with considerable knowledge of the economic doctrines of the 18th and early 19th centuries, and a highly esteemed historian who wrote a pioneering study on the Principality of Warsaw (a short-lived state created by Napoleon) based on census material. His surviving papers are in the archives of the Polish Academy of Sciences.

Selected Works

1914. *Österreichs Handelspolitik mit Bezug auf Galizien in der Reformperiode* 1772–1790. Vienna: Konegen.
1924. *Simonde de Sismondi et ses théories économiques*. Warsaw: Bibliothèque Universitaire Libre Polonaise.
1925. *Struktura spoleczna i gospodarcza Ksiestwa Warszawskiego na podstawie spisów ludności 1808–1810* (Social and economic structure of the Warsaw Principality on the basis of the census of population 1808 to 1810). Warsaw: Kwartalnik Statystyczny.
1929. *Das Akkumulations- und Zusammenbruchsgesetz des kapitalistischen Systems*. Leipzig: C.L. Hirschfeld.
1975. *Marx, l'économie politique classique et le problème de la dynamique*. Preface by Paul Mattick. Paris: Editions Champ Libre.

Grotius (de Groot), Hugo (1583–1645)

P. G. Stein

Keywords
Combinations; Contract theory; Grotius (De Groot), H.; Natural law; Private property; Usury

JEL Classifications
B31

Legal theorist, philosopher and theologian, Grotius was born in Delft on 10 April 1583 and died in Rostock on 28 August 1645. An infant prodigy, Grotius entered the University of Leyden at the age of 11, and at 15 was hailed by Henry IV of France as 'the miracle of Holland'. Deciding on a legal career, he had become Advocate General of Holland, Zealand and West Friesland by the age of 24. In this period he wrote a treatise on the Law of Prize, of which the part dealing with freedom of the seas (*Mare Liberum*) was published in 1609. Because of his support for the moderate Arminians against the Calvinists, he was in 1619 imprisoned in Loevestein castle, and while there wrote an introduction to the law of Holland (*Inleidinge tot de Hollandsche Rechtsgeleertheyd,* published in 1631) and a tract on the truth of the Christian religion, the first of many theological writings. After two years, his wife arranged his escape in a chest ostensibly holding books, and thereafter he lived mainly in France, where he served for ten years as ambassador of Sweden.

Grotius' greatest work is *De iure belli ac pacis* (On the law of war and peace), published in 1625 and widely translated (six editions appeared in English before 1750). Written during the upheavals of the Thirty Years War, it laid down certain fundamental principles of law which purported to have the certainty of mathematics and absolute validity in all times and in all places. These principles both provided a standard for measuring the validity of the positive law of any state and also formed the basis for governing the relations between states. The work had enormous influence on the ethical and legal thought of the 17th and 18th centuries and is regarded as the beginning of 'the law of nature and of nations', the forerunner of modern international law.

Grotius built on the learning of late scholastic writers on natural law, such as Suarez, but he tried to make it independent of theological doctrine, so that amid the factionalism of the Reformation its principles would be unaffected by conflicting religious views. For him these principles could be proved in two ways: a priori, by logically deducing them from the rational and social nature shared by all mankind, and a posteriori, by showing that they were generally accepted by the consensus of writers – at least in more civilized nations – through the ages. For when many writers 'at different times and different places affirm the same thing as certain, that ought to be referred to a universal cause', which must be either correct conclusion drawn from the principles of nature or common consent (Prolegomena, sec. 40). Grotius concentrated on the latter approach, and dealt particularly with property and contract, the area of law of most

concern to market societies and to nation states dealing with each other at arm's length.

Relying on the Bible narrative and on accounts of American Indians, he envisaged a primitive state of nature in which everything was held in common. When primitive simplicity gave way to specialization in agriculture and cattle raising, the conflicts that arose led first to division of lands among nations and then to division among families; thus community of property was replaced by private property. 'This happened not by a mere act of will ... but rather by a kind of agreement, either expressed, as by a division, or implied, as by occupation' (*De iure belli* II. 2, 21–5).

His doctrine of contracts was loosely based on Aristotle. He tolerated monopolies but condemned combinations to raise prices or to prevent the movement of goods by fraud or force. Although the law of nature did not forbid usury, divine positive law forbade it for Christians. However, Grotius adopted the canonist distinction between usury, which was forbidden, and receiving interest, which was permissible, if the rate was reasonable as in the positive law of Holland.

Bibliography

The most convenient modern edition of De *iure belli ac pacis* is in Classics of International Law Series, Oxford, 1925 (Vol. 1: the Latin text of 1646; Vol. 2: English translation by F.W. Kelsey). General surveys are in W.S.M. Knight, *Life and works of Hugo Grotius.* London: Sweet & Maxwell, 1925, and E. Dumbauld, *Life and writings of Hugo Grotius.* Norman: University of Oklahoma Press, 1969. Full bibliographies are in the annual volumes of *Grotiana* (New Series), Assen, Netherlands, 1980 onwards.

Group (Lie Group) Theory

Ryuzo Sato

Although it was nearly a century ago that a Norwegian mathematician by the name of Sophus Lie developed his theory of transformation groups, economists have only recently discovered that his group theory can be productively applied to such areas of economic inquiry as the theory of technical change, the theory of duality, dynamic symmetries, economic conservation laws and the theory of invariant index numbers, to name a few (see, e.g., Sato 1981). The main feature of Lie's work on transformation groups (see Lie 1888–1893 and Lie 1891; Lie and Scheffers 1893) is the study of the relationship between groups and differential equations. A survey of this particular aspect of Lie's theory is contained in the Appendix to Sato (1981).

The adoption of Lie group theory follows a long-standing tradition where economists have adapted such mathematical tools as calculus, matrix algebra, set theory, topology, probability theory, optimal control theory, game theory, etc. to their study of economic behaviour. Of course, with such a large inventory of mathematical tools, the question naturally arises: Why add group theory to the tool-kit? The answer to that question is that the application of Lie group theory is the most powerful and most systematic way of analysing 'invariant' relationships among economic variables, where often the relationships are represented by (partial) differential equation systems.

Lie Group Conception of Technical Change

To illustrate what is meant by a Lie group, consider a situation where the technical progress taking place in some production process is *a priori* known to have the simple 'neutral', or uniform factor augmenting form:

$$T_t : \overline{K} = e^{\alpha t} K, \qquad \overline{L} = e^{\alpha t} L,$$

where K represents capital, L represents labour, α ($\alpha \geq 0$) depicts the rate of technical progress, $\overline{K}$ represents 'effective' capital, $\overline{L}$ represents 'effective' labour, and t serves as the index of technical change. The equations characterizing $\overline{K}$ and $\overline{L}$ may be called the technical progress functions for capital and labour respectively.

Let the index of technical progress change from t_0 to t_1. Then, effective capital, $\overline{K}$, and effective labour, $\overline{L}$, is transformed from

$$T_{t_0} : \overline{K}_0 = e^{\alpha t_0} K, \quad \overline{L}_0 = e^{\alpha t_0} L,$$

to

$$T_{t_1} : \overline{K}_1 = e^{\alpha t_1} K, \quad \overline{L}_1 = e^{\alpha t_1} L.$$

The transformation of $\overline{K}$ and $\overline{L}$ satisfy the following conditions: (i) (*Composition*) The result of the successive performance of T_{t0} and T_{t1} is the same as that of the single transformation

$$T_{t_2} : \overline{K}_2 = \exp(\alpha(t_0 + t_1))K, \quad \overline{L}_2 = \exp(\alpha(t_0 + t_1))L.$$

(ii) (*Identity*) When there is no technical change, i.e. $t = 0$, then $\overline{K} = K$ and $\overline{L} = L$. (iii) (*Inverse*) The inverse functions of T_t are also members of T when t is replaced by $-t$, i.e.

$$T_t^{-1} = T_{-t} : K = e^{-\alpha t}\overline{K}, \quad L = e^{-\alpha t}\overline{L}.,$$

Since the transformations governing $\overline{K}$ and $\overline{L}$ satisfy the above-mentioned conditions, the technical progress functions for $\overline{K}$ and $\overline{L}$ constitute a Lie group. More specifically, they constitute a one-parameter Lie group of continuous transformations. For a more formal definition of Lie group, refer to the Appendix of Sato (1981).

Homothetic Production Functions and Hicks Neutral Technical Change

Suppose the production function characterizing the production process mentioned in the previous section was a homothetic one:

$$Y = F[f(K,L)], \tag{1}$$

where $Y =$ output, $K =$ capital, $L =$ labour, f is a continuously differentiable function homogeneous of degree one with respect to K and L, with $f_K, f_L > 0$, and $f_{KK}, f_{LL} < 0$, and F any strictly monotone increasing (or homothetic) function of f. Under the uniform factor augmenting type of technical progress mentioned in the previous section, the impact of technical progress can always be represented by another member of the class of homothetic production functions:

$$\begin{aligned} \overline{Y} &= F\left[f\left(\overline{K},\overline{L}\right)\right] = F[f(e^{\alpha t}K, e^{\alpha t}L)] \\ &= F[e^{\alpha t} f(K,L)] = G_{(t)}[f(K,L)]. \end{aligned} \tag{2}$$

The impact of this type of technical change on the technology of production is 'neutral' in the sense of Hicks. After technical change has occurred, the underlying isoquant map is *invariant*, with the exception of a change in the output levels associated with each isoquant.

Since the homotheticity property of production functions is associated with the notion of scale effects in production, the result mentioned above is disconcerting from an empirical perspective. The result implies that if time series data is used to estimate the (homothetic) form of the production function, or the rate of Hicks neutral technical change, one would not be able to distinguish between the effects of Hicks neutral technical change and returns to scale.

Holotheticity of a Technology

Consider the general technical progress functions:

$$T_t : \overline{K} = \Phi(K,L,t), \quad \overline{L} = \Psi(K,L,t). \tag{3}$$

Assume that the technical progress functions Φ and Ψ satisfy the conditions of Lie transformation groups. When technical progress occurs, it will in general affect the manner in which nominal K and L are combined in production. This in turn results in an efficiency improvement leading to effective $\overline{K}$ and effective $\overline{L}$. While there is no compelling reason for technical change to affect technology in such a way as *not* to alter the underlying isoquant map, let us confine ourselves only to the study of 'isoquant invariant' technical change.

For this kind of analysis, we require an appropriate definition.

Definition: (Holotheticity). When the complete effect of technical progress T, working through the technical progress functions Φ and Ψ within a production function f (K, L), is represented by some strictly monotone transformation F, then the production function is said to be 'holothetic'

(complete-transformation type) under a given T, i.e.:

$$\begin{aligned}\overline{Y} &= f(K,L,t) = f(\Phi,\Psi,0) = f(\Phi,\Psi)\\ &= f\left(\overline{K},\overline{L}\right) = f[\Phi(K,L,t),\ \ \Psi(K,L,t)]\\ &= g[h(K,L),t] = F_{(t)}[f(K,L)]\\ &= F_{(t)}(Y),\end{aligned} \tag{4}$$

where $h(K, L) = f(K, L, 0) = f(K, L)$.

The holotheticity condition (4) can be thought of as the condition for 'generalized' Hicks neutrality, since it encompasses all technical change (of the Lie group type), working through some production function form (technology) which would leave the underlying isoquant map invariant. Thus, we can alternatively think of the holotheticity condition as the condition for 'isoquant invariant' technical change. The impact of this kind of technical change on. the underlying production function is completely of a scale effect nature (homotheticity).

One practical value of this kind of analytical framework is that in doing empirical estimation, the researcher can determine if the hypothesized production function is holothetic under the hypothesized technical progress functions. If it is, then there is no way scale effects can be separated from the effects of technical change. If it is *not*, then scale effects can in principle be separated from the effects of technical progress. The holotheticity concept can be extended to r-parameter transformations via the concept of 'G(group) neutrality'.

Infinitesimal Operator (Lie Differentiation)

The study of production function forms (technology) which are holothetic under specific type(s) of technical progress functions is facilitated by the use of the concept of the *infinitesimal* transformation (Lie differentiation) of the technical progress functions which satisfy the Lie group properties. The infinitesimal transformations of the general technical progress functions Φ and Ψ are defined as:

$$\left.\begin{aligned}\left(\frac{\partial\overline{K}}{\partial t}\right)_{t=0} &= \left(\frac{\partial\Phi}{\partial t}\right)_{t=0} = \xi(K,L)\\ \left(\frac{\partial\overline{L}}{\partial t}\right)_{t=0} &= \left(\frac{\partial\Psi}{\partial t}\right)_{t=0} = \eta(K,L)\end{aligned}\right\} \tag{5}$$

which are components of the infinitesimal operator U:

$$U = \xi(K,L)\frac{\partial}{\partial k} + \eta(K,L)\frac{\partial}{\partial L}. \tag{6}$$

(For a more extended discussion of *Lie derivative*, see Lovelock and Ruud 1975.) By making use of (6) we can express the holotheticity condition (4) as:

$$\begin{aligned}\left(\frac{\partial Y}{\partial t}\right)_{t=0} &= Uf\\ &= \xi(K,L)\frac{\partial f}{\partial k} + \eta(K,L)\frac{\partial f}{\partial L}\\ &= G(f).\end{aligned} \tag{4'}$$

Equation (4′) points out that the holotheticity condition (4) is equivalent to the *invariance* condition of Lie group transformations.

The expression for the infinitesimal operator stated as holotheticity condition (4′) can be given the following economic interpretation:

$$\begin{pmatrix}\text{The measure of}\\ \text{technical change}\end{pmatrix} = \begin{pmatrix}\text{infinitesimal}\\ \text{transformation}\\ \text{of capital}\end{pmatrix} \times \begin{pmatrix}\text{marginal product}\\ \text{of capital}\end{pmatrix}$$

$$+ \begin{pmatrix}\text{infinitesimal}\\ \text{transformation}\\ \text{of labour}\end{pmatrix} \times \begin{pmatrix}\text{marginal product}\\ \text{of labour}\end{pmatrix} = \begin{pmatrix}\text{some transformation}\\ \text{of the production}\\ \text{function itself}\end{pmatrix}$$

Other Applications of Lie Group Theory

In the above discussion, it is demonstrated that Hicks neutral technical change (isoquant invariant) can be 'generalized' to the notion of holothetic technology.

Other areas of economic analysis that can benefit from the application of Lie group theory are comparative statics, implicit economic functions and the duality of consumer preferences and producer technologies. The fundamental theorem of comparative statics can be expressed in terms of the infinitesimal operator. It can be demonstrated that the integrability conditions of demand and utility analysis are simultaneous invariance conditions of Lie groups. Implicit economic functions and the duality and self-duality of economic functions are also areas amenable to Lie group application. For example, Lie group theory helps one to formulate exact conditions for duality and self-duality.

A theorem by Noether (1918) is most useful in the study of the invariance properties of optimal dynamic economic models, such as the Ramsey and general von Neumann types. Noether's theorem implies that if the fundamental integral of the calculus of variations is *invariant* under the *r*-parameter (Lie) group of transformations, then there are *r* conservation laws.

The theory of index numbers will also benefit from the application of Lie group theory. The Fisher-Frisch criteria of index numbers, as well as Divisia index analysis, can be represented in terms of Lie group transformations. Noether's theorem can also be applied in this area of analysis. Unquestionably, there will be many other areas of economic analysis that will benefit from the application of Lie group theory. Those readers who would like to explore the possibilities are referred to Sato (1981).

See Also

- ▶ Ces Production Function
- ▶ Meaningfulness and Invariance
- ▶ Measurement, Theory of
- ▶ Transformations and Invariance

Bibliography

Lie, S. 1888–93. *Theorie der Transformationsgruppen*. Ed. F. Engel, 3 vols, Leipzig: Teubner.

Lie, S. 1891. In *Vorlesungen über Differentialgleichungen, mit bekannten infinitesimalen Transformationen*, ed. G. Scheffers. Leipzig: Teubner.

Lie, S., and G. Scheffers. 1893. *Vorlesungen über continuierliche Gruppen mit geometrischen und anderen Anwendungen*. Leipzig: Trubner.

Lovelock, D., and H. Ruud. 1975. *Tensors, differential forms, and variational principles*. New York: Wiley.

Noether, E. 1918. Invariante variationsprobleme. *Nachrichten Akademie Wissenschaft Gottingen, Mathematischen-Physischen Kl.II*. Trans. M.A. Tavel as 'Invariant variational problems', *Transport Theory and Statistical Physics* 1(3).

Sato, R. 1975. *The impact of technical change on the holotheticity of production functions*. Paper presented at the World Congress of the Econometric Society, Toronto. Published as Sato (1980).

Sato, R. 1980. The impact of technical progress on the holotheticity of production functions. *Review of Economic Studies* 47: 767–76.

Sato, R. 1981. *Theory of technical change and economic invariance: Application of Lie Groups*. New York: Academic Press.

G

Group Selection

Arthur J. Robson

Abstract

'Group selection' is the biological term for the possibility that a characteristic that is beneficial to the group but possibly costly to the individual is evolutionarily successful. Although logically possible, it is generally viewed with scepticism by biologists. It is not problematic that group selection would favour an equilibrium whose payoffs dominated those of another, because there is then no conflict with individual selection within each group. Group selection might reject inefficient equilibria in a repeated game, for example. Since human societies can support rather arbitrary outcomes as equilibria, group selection could play a role in human evolution.

Keywords

Altruism; Assortative matching; Cooperation and its evolution; Cultural transmission; Darwin, C.; Evolution; Exchange; Group selection; Hunter–gatherer economies; Individual selection; Kin selection; Natural selection; Other-regarding preferences; Prisoner's dilemma; Punishment; Repeated games; Trust

JEL Classifications

C71; C73

It is uncanny how close Darwin (1859) came to the modern view of biological evolution, given that detailed understanding of the mechanics of genetic inheritance lay in the future. Darwin understood how mutations might arise randomly rather than in response to circumstances. Further, he espoused the modern dogma concerning the separation of the germ line (sex cells) and the somatic line (all other bodily cells). Under this dogma, which explicitly contradicts the earlier biologist Lamarck, only those characteristics present in the germ line, not those acquired during the individual's lifetime, are genetically inherited by offspring. If a germ line characteristic produces a somatic or behavioural attribute that is best suited to the ecological or social circumstances of the individual (yielding the most offspring), then the attribute and the underlying genetic characteristic will become more common.

Darwin typically emphasized that a particular variation would spread if this variation led to greater reproductive success for individuals and were inherited by their descendants. However, Darwin also strayed occasionally into what would now be called 'group selection', especially when he was considering the implications of his theory for humans (Darwin 1871, p. 166). Thus, he thought an individual human might engage in behaviour that is beneficial to the survival of a group, even if this behaviour had a fitness cost to the individual. It is of interest that Darwin refers to humans in particular, since it is still sometimes argued that group selection might be important for our own species.

The Group Selection Debate in Biology

There is a 'folk wisdom' appeal to group selection in biology. This mechanism was once invoked in popular accounts of natural selection. For example, the idea that a predator species is doing a prey species a favour by eliminating its weakest members involves an egregious form of group selection. The English experimental biologist Wynne-Edwards provided an especially explicit manifesto on group selection and became thereby a favourite target for those wishing to argue against it. In Wynne- Edwards (1962), for example, he argued that birds limit the size of their clutches of eggs to ensure that the size of the population does not exceed the comfortable carrying capacity of the environment.

These particular group selection arguments were effectively devastated by Williams (1966). If a new type of individual does not so obligingly limit her clutch, why would this more fertile type not take over the population, without regard for the standard of living? Indeed, there are compelling arguments why it is in the interests of the individual to limit her clutch size. For example, it might well be that, beyond a certain point, an increase in the number of eggs reduces the expected number of offspring surviving to maturity, because each egg then commands a reduced share in parental resources. A finite optimum for clutch size is then to be expected.

Dawkins (1976 and 1982, for example) has been even more insistent than Williams on rejecting group selection, going further in arguing for the primacy of the gene as a still lower-level unit of selection. There certainly are phenomena best understood at the level of the gene. Consider, for example, sickle-cell anemia. At the relevant locus, there exists a particular variant gene, a particular allele, that is. If one of the two alleles that are present at this locus is this variant gene, the individual has improved resistance to malaria. However, if both alleles are of this variant type, the red blood cells have a characteristic sickle shape. Such cells malfunction by not carrying adequate oxygen, implying increased mortality. Under sexual reproduction, there is no way of maintaining only the individuals who have

exactly one copy of the variant gene. Rather, both alleles are maintained as a stable mixture, where each allele is present in individuals who have quite different fitnesses.

There are presumably a fair number of cases where the interests of the gene and the individual do not conflict. In any case, it is often difficult to give concrete form to the gene as the unit of selection, given our ignorance of the details of the transformation of genes into individuals, particularly for complex behavioural characteristics (Grafen 1991, advocates finessing such detailed questions on the genetic basis of individual variation. This is his so-called 'phenotypic gambit'). Hence, despite the theoretical primacy of the gene, we restrict attention here to the comparison between the individual level and the group level of selection.

In order to fix ideas, consider the following outline of the classic model that addresses the issue of individual selection versus group selection.

The Haystack Model

The following is a simplified account of Maynard Smith (1964). There are a number of haystacks in a farmer's field, each of which is home to two mice. Each pair of mice now play the Prisoner's Dilemma game, with the usual two choices: cooperate or defect. Payoffs for each individual take the concrete form of the number of offspring, but reproduction is asexual, with offspring inheriting their mother's choice of strategy. There are a number of subsequent stages of play, where the mice in each haystack are paired at random to play the Prisoner's Dilemma game. The number of individuals within the haystack choosing each strategy then grows in an endogenous fashion, as does the overall size of the group. Every so often, once a year, say, the haystacks are removed, and the mice are pooled into a single large population. Now pairs of mice are selected at random from the overall population to re-colonize the next set of haystacks, and excess mice die.

Maynard Smith's intention here was to give the devil his due, by building a model in which group selection might well have an effect. At the same time, he wished to show that the parametric assumptions needed to make group selection comparable in strength to individual selection would be unpalatable. In order for group selection to be effective there must be a mechanism that insulates the groups from one another. Only then can a cooperative group be immune to infection by a defecting individual, and maintain its greater growth rate. Even with the temporary insulation of each haystack in this model, cooperation will evolve only if there are sufficient rounds of play within each haystack, so that defectors from a particular haystack are likely to be eliminated when the groups are reformed.

A loose description of the problem with group selection is that it relies too heavily upon a group becoming extinct as a likely consequence of a choice that is bad for the group. There is clearly scope, in reality, for individual selection, since individuals die frequently; group selection is less plausible, since there may not be enough extinction of groups.

An Example of Group Selection?

Despite the disfavour into which group selection has fallen in biology, there remain examples of cases that are challenging to explain otherwise. One of these concerns the interaction of the myxoma virus and European rabbits in Australia (Lewontin 1970, proposed a group selection interpretation of this case. Sober and Wilson 1999, pp. 45–50, give a – somewhat partisan – view of the subsequent debate). English rabbits were introduced into Australia in a misguided attempt to recreate the English countryside in Australia, but their numbers grew out of control, causing massive economic damage to farms there. The myxoma virus was first identified in South American forest rabbits, where it was only mildly virulent. When this South American strain was originally tested on Australian rabbits, however, it seemed an ideal solution to the rabbit infestation there, since it killed nearly the entire sample. Unfortunately, in the long run, after the South American virus had been present in the Australian

rabbit population for a while, the rabbit mortality rate fell dramatically.

Why? Perhaps the most obvious explanation is that the rabbit population had been selected to have greater resistance to the virus, consistent with individual selection of rabbits. That this was true to some extent was demonstrated by the finding that the new Australian strain of the virus had a greater effect on laboratory- bred Australian rabbits than on the feral stock. However, this effect was not sufficient to explain the entire drop in rabbit mortality. Indeed, both laboratory and wild strains of Australian rabbits were less susceptible to the new Australian strain of the virus than they were to the original strain imported from South America. The virus had evidently evolved to be less virulent as a result of its interaction with the Australian rabbit population.

This situation might be roughly mapped onto the haystack model as follows. Consider a group of viruses to be those infecting a given rabbit. The evolutionary success of this group might best be promoted by settling for a moderate level of mortality for the host rabbit. Whatever the other advantages to the virus of strategies that induce high mortality in the rabbit, the early death of the current host makes transmission to a new host difficult. However, prolonging the life of the host is a public good from the point of view of the infecting viruses. A strain of virus with a strategy that was more lethal to the host could then grow as a fraction of the group of viruses. As in the haystack model, however, if there were enough generations of the virus within each rabbit's lifetime, this conflict between the group and the individual might be resolved in favour of the group, as suggested by the data.

Selection Among Equilibria

When does group selection matter in biology? In theory, it could lead to different results than would individual selection, as in the Prisoner's Dilemma, and as it may in above example. In practice, the above example is atypical, and group selection is usually a rather weak force. There is, however, one compelling scenario in which group selection would operate robustly, in any species. This is as a mechanism to select among equilibria (Boyd and Richerson 1990). Consider a population that is divided into various sub-populations, which are largely segregated from one another, so that migration between sub-populations is limited. Each sub-population plays the same symmetric game, which has several symmetric equilibria. Play of this game involves a random matching of the members of each sub-population. Individual selection ensures that some equilibrium is attained, within each sub-population. But group selection is then free to operate in a leisurely fashion to select the Pareto- superior equilibrium, since there is no tension here between the two levels of selection.

Group Selection and Economics

Why does group selection matter in economics? Group selection is the most obvious mechanism for generating preferences in humans that might make them behave in the social interest rather than that of the individual. At stake, then, is nothing less than the basic nature of human beings. As an economist, one should be sceptical of the need to suppose that individuals are motivated by the common good. Economic theory has done well in explaining a wide range of phenomena on the basis of selfish preferences, and so the view of the individual as the unit of selection is highly congenial to economists. Furthermore, to the extent that armchair empiricism suggests that non-selfish motivations are sometimes present, these seem as likely to involve malice as to involve altruism. For example, humans seem sometimes motivated by relative economic outcomes, which involve such a negative concern for others. Group selection is a blunt instrument that might easily 'explain' more than is true.

There are, nevertheless, some aspects of human economic behaviour that are tempting to explain by group selection. For example, we have a proclivity for trade that may go beyond myopic self-interest. As Darwin, an astute observer of both human beings and the natural world, observed on one of his visits to Tierra del Fuego,

> Some of the Fuegians plainly showed that they had a fair notion of barter. I gave one man a large nail (a most valuable present) without making any signs for a return; but he immediately picked out two fish, and handed them up on the point of his spear. (Darwin 1845, ch. 10)

That is, human beings are often willing to trade with strangers they will likely never see again, as might be analogous to cooperating in the one-shot Prisoner's Dilemma. There is no shortage of reliable data showing that human beings are capable of such apparently irrationally cooperative behaviour, in appropriate circumstances. Whatever the underlying reasons for this, it is clearly significant, and may even help account for the prodigious economic and biological success of humans.

Furthermore, identifying the underlying reasons would help shape a theory of such behaviour that remains falsifiable; such a theory might also predict what alternative circumstances might induce non-cooperative or antagonistic behaviour. Group selection is an obvious avenue to explore in this connection.

It is sometimes argued, in particular, that the structure of hunter–gatherer societies helps account for cooperative behaviour. Hunter–gatherer societies were composed of a large number of relatively small groups, and individuals within each group were often genetically related. Perhaps, so the argument goes, we acquired an inherited psychological inclination towards conditional cooperation in such a setting, partly perhaps as a result of group selection. These inclinations then carried over into modern societies, despite genetic relatedness now being essentially zero on average. Seabright (2004), for example, argues eloquently that human societies cannot function on myopic self-interest alone, but also that the trust needed for exchange in market economies sprang from adaptations to archaic small groups. It is hard to believe, however, that hunter–gatherers never encountered strangers. If there were good reasons to condition on this distinction, why would corresponding different strategies not have evolved?

A phenomenon that looms large in the case of human beings is culture, by which is meant the non-genetic transmission of behaviour, by imitation of peers, for example. Many attempts to derive cooperative behaviour have focused then on group selection in models of cultural transmission. We now turn to this literature.

Cultural Group Selection and Economics

A spectacular and famous example of cultural group selection features the Nuer and Dinka, who lived as neighbouring ethnic groups in 18th century southern Sudan (Kelly 1985). Despite the similarity of their environment, these two groups differed in various economic and political respects. Perhaps the key difference was that Dinka lived in small groups, the size of which was related to the needs of their economic activity. The Nuer, on the other hand, organized their society according to a patrilineal system that bound many such smaller units together, over a greater geographic area. Over a period of 100 years, the Nuer expanded their territory at the expense of the Dinka, killing or assimilating their rivals. Nuer culture, that is, was selected over that of the Dinka.

Despite the apparently strong military advantages of the Nuer political system over that of the Dinka, it seems plausible that any individual – Dinka or Nuer – would have had the incentive to play the appropriate role within their society. It would not have been possible for an individual Dinka to shift unilaterally to Nuer behaviour.

Human societies have the capacity to render a wide range of behaviour optimal for the individual. If a society wishes to adopt a rather arbitrary rule, it may well have adequate sanctions to enforce this (Boyd and Richerson 1992). As described above, group selection can then be relied upon to select between various equilibria. Boyd et al. (2003) make the important additional point that, because punishment is needed only infrequently near full cooperation, the weak force of group selection can support cooperation as an equilibrium, without the usual need for punishment of those who fail to punish, and so on.

Group selection is uncontroversial as a mechanism for selecting an efficient outcome within each group. But this does not directly explain

observations, such as Darwin's, of our apparent willingness to trade with strangers. The difficulty is stark: defection is a strictly dominant strategy in the one-shot Prisoner's Dilemma game.

One stark option is that cooperation is hard-wired. Bowles et al. (2003) argue that, if behaviour is directly genetically controlled, cooperative behaviour may then be sustained in the presence of culturally maintained institutions. These institutions, food-sharing for example, serve to reduce the negative impact on individuals of cooperative behaviour. Group selection arises from conflict between groups, with more cooperative groups emerging victorious in such conflict.

However, human strategic behaviour is genetically mediated in a complex and poorly understood way, and is not always cooperative. Even if we did somehow acquire a genetic inclination to cooperate in archaic societies, shouldn't we now be in the process of losing this inclination in modern large and anonymous societies?

A Recent Revival?

Sober and Wilson (1999) push energetically for a rehabilitation of group selection within biology. They argue that group selection is closely related to other well- accepted phenomena. For example, they argue that kin selection – the widely accepted notion that individuals are selected to favour their relations – should be regarded as a special case of group selection.

In its simplest form, kin selection is the argument that individuals should undertake actions that benefit a relation if this benefit, when deflated by the degree of relatedness, exceeds the cost to the first individual (This is 'Hamilton's rule' as in Hamilton 1964). The empirical evidence in favour of kin selection is overwhelming mothers, human and non-human, routinely make large sacrifices in favour of offspring. Even economists would exempt such interactions from the general presumption of selfish behaviour.

Sober and Wilson certainly make the case that these phenomena can be viewed in an integrated fashion. Indeed, it is a consequence of the 'Price equation' (Price 1970) that what matters most fundamentally is the likelihood that altruistic individuals will be preferentially matched with other altruistic individuals. In the case of kin selection, this could be ensured by a mechanism to directly recognize relations by smell, for example – or by indirect but reliable methods, such as, for example, assuming that those who are found in proximity to one's mother are relations.

Bergstrom (2002) provides the best introduction to the literature on group selection for economists. He presents a unified and intuitive treatment of the literature, and also stresses the key role of assortative matching. Thus, for example, if the subgroups in the haystack model are dispersed after one round of the game, and then randomly recombined, there is no force to group selection. Only if the subgroups remain together for repeated play of the game is there effective assortative matching. Such a structure seems less compelling for non-relations than it is for relations.

Despite the formal analogies between kin selection and group selection, acceptance of the former does not then require acceptance of the latter. In the end, a sceptical but not dogmatic view of the importance of group selection to human economic behavior seems warranted.

See Also

- ▶ Game Theory and Biology
- ▶ Hunting and Gathering Economies
- ▶ Learning and Evolution in Games: An Overview

Acknowledgment I received helpful comments from Ted Bergstrom, Lawrence Blume, Sam Bowles, Steven Durlauf and Peter Sozou. I thank them without blaming them.

Bibliography

Bergstrom, T.C. 2002. Evolution of social behavior: Individual and group selection. *Journal of Economic Perspectives* 16(2): 67–88.

Bowles, S., J.-K. Choi, and A. Hopfensitz. 2003. The coevolution of individual behaviors and social institutions. *Journal of Theoretical Biology* 223: 135–147.
Boyd, R., and P. Richerson. 1990. Group selection among alternative evolutionarily stable strategies. *Journal of Theoretical Biology* 145: 331–342.
Boyd, R., and P. Richerson. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* 13, 171–195.
Boyd, R., H. Gintis, S. Bowles, and P.J. Richerson. 2003. The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences* 100: 3531–3535.
Darwin, C. 1845. *Journal of researches into the natural history and geology of the various countries visited by H. M. S. Beagle.* London: Henry Colburn. Online. Available at http://www.galapagos.to/TEXTS/J-OF-R.HTM. Accessed 23 Oct 2006.
Darwin, C. 1859. *On the origin of species by means of natural selection.* London: John Murray. Online. Available at http://darwin-online.org.uk/EditorialIntroductions/Freeman_OntheOriginofSpecies.html. Accessed 8 Nov 2006.
Darwin, C. 1871. *The descent of man and selection in relation to sex.* London: John Murray. Online. Available at http://darwin-online.org.uk/EditorialIntroductions/Freeman_TheDescentofMan.html. Accessed 8 Nov 2006.
Dawkins, R. 1976. *The selfish gene*. Oxford: Oxford University Press.
Dawkins, R. 1982. *The extended phenotype: The gene as the unit of selection*. Oxford: Oxford University Press.
Grafen, A. 1991. Modelling in behavioural ecology. In *Behavioral ecology: An evolutionary approach*, 3 ed., ed. J.R. Krebs and N.B. Davies. Oxford: Blackwell Scientific.
Hamilton, W.D. 1964. The genetical evolution of social behavior. *Journal of Theoretical Biology* 7: 1–16.
Kelly, R.C. *1985. The Nuer conquest: The structure and development of an expansionist system*. Ann Arbor: University of Michigan Press.
Lewontin, R.C. 1970. The units of selection. *Annual Reviews of Ecology and Systematics* 1: 1–18.
Maynard Smith, J. 1964. Group selection and kin selection. *Nature* 201: 1145–1147.
Price, G. 1970. Selection and covariance. *Nature* 277: 520–521.
Seabright, P. 2004. *The company of strangers*. Princeton: Princeton University Press.
Sober, E., and D.S. Wilson. 1999. *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.
Williams, G.C. *1966. Adaptation and natural selection: A critique of some current evolutionary thought*. Princeton: Princeton University Press.
Wynne-Edwards, V.C. 1962. *Animal dispersion in relation to social behavior*. Edinburgh: Oliver and Boyd.

Growth Accounting

Francesco Caselli

Abstract

Growth accounting consists of a set of calculations resulting in a measure of output growth, a measure of input growth, and their difference, most commonly referred to as total factor productivity (TFP) growth. It can be performed at the level of the plant, firm, industry, or aggregate economy. This article discusses the theoretical interpretation of the growth-accounting exercise, problems of measurement, and main empirical results. It concludes with a (very selective) history of the field.

Keywords

Growth accounting; Input–output analysis; Measurement error; National accounts; Production functions; Research and development; Specification error; Technical change; Total factor productivity; Vintage capital

JEL Classifications

D4; D10

Growth accounting consists of a set of calculations resulting in a measure of output growth, a measure of input growth, and their difference, most commonly referred to as total factor productivity (TFP) growth. It can be performed at the level of the plant, firm, industry, or aggregate economy.

Current growth-accounting practice tends to rely on the theoretical construct of the production function both as a guide for measurement and as a source of interpretation of the results. Apart from the existence of a production function linking inputs and outputs, the main assumption is that factors of production are rewarded by their marginal product. In continuous time, this permits a

representation of output growth as a weighted sum of the growth rates of the inputs, and an additional term that captures shifts over time in the production function. The weights for the input growth rates are the respective shares in total input payments. Since data on the growth of output and individual input quantities cover discrete periods of time, a discrete-time approximation to the weights is required. Current practice tends to use simple averages of the input shares at the beginning and the end of each period. In the special case that the production function is of the translogarithmic form, this procedure actually results in an exact decomposition; otherwise it can be interpreted as a second-order approximation.

It is customary to group inputs into broad categories. When output is measured as value-added the broad categories are labour and capital. When output is total production one has to add materials, which are occasionally further broken down with further entries for energy and services (giving rise to the so-called KLEMS accounting framework). This kind of grouping allows one to speak of, for example, the 'contribution of labour (capital, materials) to output growth'. However, this grouping masks an enormous heterogeneity of the underlying inputs. This heterogeneity is the source of a large share of the measurement problems in growth accounting. These problems are most severe in the measurement of the growth of capital input.

Capital inputs are heterogeneous within vintages (for example, tractors versus personal computers) and across vintages (computers produced in 2006 versus computers produced in 2007). Heterogeneity within vintages is best addressed by having as fine a disaggregation of capital types as the data will allow. The most important data constraint on disaggregation of capital types occurs in the construction of type-specific shares in total capital income, as these require type-specific estimates of rental rates, which in turn require type-specific estimates of depreciation rates, capital gains and tax treatment. Heterogeneity across vintages, also known as embodied technical change, or quality change, poses even more difficult problems. Most practitioners' ideal solution to this problem is to put the measurement of the stocks of different types of capital on a constant-quality basis, by applying appropriate deflators reflecting quality change to the corresponding investment series. However, the availability and/or accuracy of such deflators, whose construction generally requires hedonic methods, is currently limited for most countries, industries and capital types. As a result there is a presumption that the growth rate of (the efficiency units embodied in) the capital stock is often understated.

Construction of indices of labour input growth have conceptually similar problems. However, aggregation across types (for example, female, white, highschool graduates, of age 40–45 versus male, black, college graduate of age 35–40) is simpler as average rental rates (that is, hourly wages) for reasonably fine categories are reasonably well observed (while in the case of capital goods they must be estimated). The vintage problem is typically bypassed by assuming that there is no quality change within narrowly defined categories.

Another difficult problem is how to turn the growth in input stocks into growth in the flow of input services, that is, how to account for variation in the rate of utilization of labour and capital. Measuring labour in hours is helpful, but an issue of utilization still remains if effort per hour is not constant, as is likely. For capital, various adjustments based on proxies for utilization have been proposed, a classic one being a measure of electricity consumption. But this approach to the problem of measuring utilization creates a deeper problem of interpretation, or at least a conflict with the estimate of rental rates. This is because the latter are constructed in a way that assumes them to be invariant to the rate of utilization. But in this case the opportunity cost of setting the utilization rate to 100 per cent all the time is nil, and there should be no variation in utilization. Some more systematic adjustment to the theoretical framework, such as endogenous depreciation or limited opportunity for substitution between capital and other inputs, is therefore required to fully solve the measurement and interpretation challenges posed by variable utilization.

At the plant, firm, and industry levels a choice can be made between accounting for total

production or value added. The total production approach is attractive, because after all it is total production that 'comes out' of the production process. Furthermore, the conditions for existence of a well-defined production function for total output are far less stringent than the conditions for existence of a function linking valued added to capital and labour inputs. On the other hand, the results of growth-accounting exercises based on total output are very sensitive to the degree of vertical integration, and this causes severe problems of interpretation.

At the country level value added is obviously the only meaningful concept of output, no matter how stringent the conditions for an aggregate valued-added function. Because of the well-known shortcomings of standard measures of valued added as indicators of the 'want-satisfying' capacity of the economy, some attempts have been made to augment such measures by estimates of non-market outputs, chiefly the output of the education sector. Accounting for the effects of economic activity on the natural environment is very probably the next frontier.

Among the outputs of the growth-accounting calculation the one to receive most attention is usually the difference between output and input growth. This is somewhat surprising because the interpretation of this quantity is fraught with difficulties, as underscored by the multitude of phrases used to refer to this difference: besides TFP growth, 'multi-factor productivity' growth, '(Solow) residual', 'measure of our ignorance', 'rate of technical change', and growth in 'output per unit of (total) input', among others.

What is sometimes misunderstood is the relationship between the difference and technical change. An economic unit can use additions to its capital and labour either to directly produce more output, or to devise ways to rearrange the existing capital and labour so as to produce more (constant-quality units of) output, the latter being the definition of research and development (R&D). If it does so by equating the marginal products of labour and capital between direct production of output and indirect production of output through R&D, the extra output produced thanks to R&D will be fully 'accounted for' by the measured growth in capital and labour inputs. Hence, TFP growth does not really measure technical change as this term is commonly understood. Furthermore, failures of TFP growth to accelerate in periods/ industries/firms experiencing increases in R&D spending do not need to be puzzling.

For the same reasons, TFP growth can be identified neither with disembodied nor with embodied technical change. Embodied technical change in capital-using industries is a reflection of disembodied technical change in capital-producing industries, but neither need necessarily show up in the TFP numbers, as long as R&D costs have been properly accounted for.

So what does show up? Under the maintained theoretical assumptions, the cleanest interpretation –apart from weather shocks, and costless, instantaneous flashes of inspiration (if they were not instantaneous an opportunity cost of time would have to be imputed) or innovations stumbled upon by luck, none of which seems susceptible to vary much over time and space, or with government policy – is R&D externalities. If the units performing R&D fail to capture all the social return from it, other units will experience costless growth in output per input, and this will be detected by TFP growth. Under this interpretation, a link may indeed be found between R&D and TFP growth, and if so it would be possible to use the framework to advocate policies to encourage R&D. Other forms of externalities may also give rise to positive TFP growth.

But since TFP is a residual, it also picks up, as all residuals, errors of specification and measurement. We have already discussed mismeasured input growth, chiefly in terms of incomplete adjustment for quality change. A failure to account for quality change in output will push TFP growth in the opposite direction. Note that mismeasured quality change in capital results in lower TFP growth in capital-producing industries, higher TFP growth in capital-using industries, and somewhat ambiguous effects on TFP at the aggregate level, though the net effect is usually deemed to be positive.

Many economies are likely to be characterized by frictions to the efficient allocation of resources

among economic units, implying that marginal products of homogeneous inputs are not equalized. In these cases improvements in the allocation of resources will also result in positive TFP growth.

It is impossible to overestimate the interest that growth-accounting calculations have elicited. There must be very few industries and countries for which some kind of input–output data exists that has not been used for performing a growth accounting exercise. Indeed, several national statistical agencies explicitly include the output of growth-accounting calculations, including TFP growth, into the national accounts.

I am unable to provide here an overview of this immense body of work, and the reader will have to refer to the country/industry/period of interest on a case-by-case basis. However there are a couple of broad lessons that can be distilled here. First, over the medium to long term, the residual accounts for a relatively minor portion of overall growth in output. For example for the United States it is possible to explain about two-thirds of growth in (market) output per worker over the post-war period by changes in the quality and quantity of inputs. For countries experiencing exceptionally high growth rates, such as the Asian Tigers between 1960 and 1990, this share is even higher. To the extent that the residual picks up measurement and specification errors, this is tantamount to saying that the performance of the growth accounting methodology is very good by the standards of empirical work in economics. This interpretation is reinforced by the fact that, again by and large, the role of the residual tends to be systematically smaller in studies deploying better quality data.

Over shorter horizons, however, TFP growth is harder to underplay. For example a slowdown in TFP growth 'accounts' for a large fraction of the slowdown in output growth observed between the mid-1970s and the mid-1990s. Not coincidentally, the root causes of that slowdown remain as mysterious as ever.

While growth-accounting calculations can be performed at various levels of aggregation, and their interpretation is perhaps easier the smaller the unit of analysis, the origins of growth accounting are macroeconomic. The earliest growth-accounting exercises (Stigler 1947; Schmookler 1952; Abramowitz 1956 – the latter also coining the expression 'measure of our ignorance') were a direct byproduct of the development of US aggregate national account data. One exception was agriculture, for which early growth-accounting experiments date to 1948 (Barton and Cooper) and 1951 (Kendrick and Jones). Kendrick (1956, 1961) compiled the first large-scale growth-accounting calculations broken down by many industries. He also introduced the phrase 'total factor productivity'.

Solow (1957) laid out the theoretical foundations of growth accounting (a previous contribution in this direction by Tinbergen 1942, with attendant calculations, was discovered by the English-language literature only subsequently). Solow (1960) and Jorgenson (1966) worked out the implications of embodied technical change. Denison (1962) introduced corrections for changes in the composition of the labour force. Griliches and Jorgenson (1966) and Jorgenson and Griliches (1967) put aggregation of inputs and outputs on a solid theoretical basis, particularly by showing how to correctly estimate rental rates. They also pioneered empirical approaches to quality change and variable utilitization. This programme was further refined by Christensen and Jorgenson (1969, 1970) for the aggregate economy, and Fraumeni et al. 1987) for a broad set of industry-level calculations which has shaped the way US national accounts are now constructed, and whose methods are widely accepted to be the gold standard for the purposes of productivity measurement.

Christensen et al. (1973) developed the translogarithmic production frontier, and Diewert (1976) showed that with translog production functions discrete-time approximations are no longer approximations. Jorgenson and Fraumeni (for example, 1992) attempted accounting for the output of the education sector. Young (1995) performed an influential growth-accounting exercise for the East Asian tigers.

See Also

► Level Accounting

Bibliography

Abramowitz, M. 1956. Resource and output trends in the United States since 1870. *American Economic Review* 46: 5–23.

Barton, G., and M. Cooper. 1948. Relation of agricultural production to inputs. *The Review of Economics and Statistics* 30: 117–126.

Christensen, L., and D. Jorgenson. 1969. The measurement of U.S. real capital input, 1929–1967. *Review of Income and Wealth* 15: 293–320.

Christensen, L., and D. Jorgenson. 1970. U.S. real product and real factor input, 1929–1967. *Review of Income and Wealth* 16: 19–50.

Christensen, L., D. Jorgenson, and L. Lau. 1973. Trascendental logarithmic production frontiers. *The Review of Economics and Statistics* 55: 28–45.

Denison, E. 1962. *The sources of economic growth in the United States and the alternatives before Us*. New York: Committee for Economic Development.

Diewert, E. 1976. Exact superlative index numbers. *Journal of Econometrics* 4: 115–145.

Fraumeni, B., F. Gollop, and D. Jorgenson. 1987. *Productivity and U.S. economic growth*. Cambridge, MA: Harvard University Press.

Griliches, Z., and D. Jorgenson. 1966. Sources of measured productivity change: Capital input. *American Economic Review* 56: 50–61.

Jorgenson, D. 1966. The embodiment hypothesis. *Journal of Political Economy* 74: 1–17.

Jorgenson, D., and B. Fraumeni. 1992. Investment in education and U.S. economic growth. *Scandinavian Journal of Economics* 94 (Suppl): 51–70.

Jorgenson, D., and Z. Griliches. 1967. The explanation for productivity change. *Review of Economic Studies* 34: 249–283.

Kendrick, J. 1956. Productivity trends: Capital and labor. *The Review of Economics and Statistics* 38: 248–257.

Kendrick, J. 1961. *Productivity trends in the United States*. Princeton: Princeton University Press.

Kendrick, J., and C. Jones. 1951. Gross national farm product in constant dollars, 1910–50. *Survey of Current Business* 31: 13–19.

Schmookler, J. 1952. The changing efficiency of the American economy, 1869–1938. *The Review of Economics and Statistics* 34: 214–231.

Solow, R. 1957. Technical change and the aggregate production function. *The Review of Economics and Statistics* 39: 312–320.

Solow, R. 1960. Investment and technical progress. In *Mathematical methods in the social sciences, 1959*, ed. K.J. Arrow, S. Karlin, and P. Suppes. Stanford: Stanford University Press.

Stigler, G. 1947. *Trends in output and employment*. Cambridge, MA: NBER.

Tinbergen, J. 1942. Zur Theorie der Langfristigen Wirtschaftsentwicklung. *Weltwirtschaftliches Archiv* 55: 511–549.

Young, A. 1995. The tyranny of numbers: Confronting the statistical realities of the East Asian growth experience. *Quarterly Journal of Economics* 110: 641–680.

Growth and Civil War

Paul Collier

Abstract

Civil war is obviously damaging for both society and the economy. The social consequences are often difficult to measure: for example, people die as a result of disease and are traumatized through rape or experience as child soldiers. However, the consequences for the economy are much more amenable to quantification, and a key economic research issue has been to try to classify and quantify the economic damage.

Keywords

Capital flight; Civil war; Economic growth; Foreign aid; Human capital; Predation; Social cost; Spillovers

JEL Classifications

O4

Civil war is obviously damaging for both society and the economy. The social consequences are often difficult to measure: for example, people die as a result of disease and are traumatized through rape or experience as child soldiers. However, the consequences for the economy are much more amenable to quantification, and a key economic research issue has been to try to classify and quantify the economic damage.

There are three consequences of civil war for economic growth. The most obvious is the loss of growth experienced by the country during the war. War is destructive of both the capital stock and normal economic activity. In addition to inflicting purposive destruction, rebel forces typically prey

on economic activity within their military reach, since they have no official sources of finance. Government forces may adopt the same tactics: typically, lines of command are too loose to prevent the diversion of military force into decentralized predation. In such circumstances citizens with assets radically reduce their investment and resort to capital flight abroad, and poorer citizens may similarly retreat into subsistence activities that are less vulnerable. The consequences of civil war for the growth of gross domestic product (GDP) have been estimated from time series regressions, a typical result being that the annual growth rate is reduced by around 2.3 percentage points (Collier 1999). Since the typical civil war lasts around seven years, by the end of the conflict the country is around 15 per cent poorer than it would have been had it remained at peace.

The second consequence for growth is the legacy of war – effects that arise after the war has ended. Peace does not usually enable the economy to rebound swiftly to its previous growth path. Even if the post-conflict peace is secure the economy will take several years to rebuild its capital stock, and some costs, such as a loss of human capital, may be irreversible. More typically, the legacy of civil war creates major problems for economic recovery. The peace itself may be insecure: around 40 per cent of post-conflict situations revert to conflict within a decade. The breakdown in social order during civil war allows opportunistic behaviour to become more prevalent. Both the high macro-risk of conflict reversion and the high micro-risk of being the victim of opportunism inhibit investment, and capital flight typically continues. These effects dampen and can potentially prevent post-conflict recovery. Often, post-conflict growth exceeds normal growth by around 1.1 per cent, so that after a war the economy needs about double the length of time of the civil war before it rejoins its long-term path (Collier and Hoeffler 2004a). Until then the country is poorer than it would have been without the war, and so this shortfall should properly be counted as a cost. Because recovery is so slow, most of the costs of a civil war accrue after it is over. Aid accelerates growth during recovery and has been found to be particularly effective in post-conflict situations. This was indeed the original rationale for aid: the World Bank was initially going to be named the 'International Bank for Reconstruction'.

The third consequence for growth is the effect of civil war on other countries. Typically, the adverse effects of a civil war spread far beyond the borders of the afflicted country: even countries that are not direct neighbours suffer significant reductions in growth (Murdoch and Sandler 2002). Although the reduction in the growth rate of neighbours is much lower than that experienced by the country itself, because many countries are affected, the overall 'spillover' costs to neighbours arising from this loss of growth tend to exceed those experienced by the country itself.

Taken together, these consequences for growth have two important implications. One is that the true economic cost of civil war is massive even before one takes into account the social costs. By applying a conventional discount rate to the lost growth and allowing only for costs to direct neighbours, Collier and Hoeffler (2004b) estimate the value of the typical cost of a civil war in a low-income country at an astounding 64 billion dollars. The other implication is that most of these losses are externalities to the people taking the decisions at the time of the conflict: the losses accrue to neighbours and to a future generation. Hence, decisions to start civil wars are unlikely to reflect a true social calculus of the probable consequences of war.

Bibliography

Collier, P. 1999. On the economic consequences of war. *Oxford Economic Papers* 51: 168–183.

Collier, P., and A. Hoeffler. 2004a. Aid, policy and growth in post-conflict societies. *European Economic Review* 48: 1125–1145.

Collier, P., and A. Hoeffler. 2004b. Conflict. In *Global crises, global solutions*, ed. B. Lomborg. Cambridge: Cambridge University Press.

Murdoch, J.C., and T. Sandler. 2002. Economic growth, civil wars and spatial spillovers. *Journal of Conflict Resolution* 46: 91–110.

Growth and Cycles

Gadi Barlevy

Abstract

There is a long tradition in macroeconomics of treating growth and cycles as distinct phenomena. However, various economists have also recognized the virtue of incorporating the two forces into a single framework and to study the way they are related. This article reviews this literature, with emphasis on attempts not only to integrate growth and cycles into a single framework but also to endogenize growth, cycles, or both.

Keywords

Capital accumulation; Endogenous growth; Growth and cycles; Innovation; Kydland, F.; Prescott, E.; Ramsey model; Research and development; Risk; Technical change; Precautionary savings

JEL Classifications

O4

Growth and cycles are two key features that characterize real output per capita in most industrialized countries. Real output per capita grows systematically over time; and the rate at which it grows tends to fluctuate over time.

A long tradition in macroeconomics treats these two features as distinct. On the one hand, economists who study why output per capita consistently grew in most countries during the 20th century often ignore the fact that growth in any given country was uneven over time. Underlying this approach is the assumption that temporary fluctuations in economic growth are transitory and have no consequences for long-run growth. On the other hand, economists interested in cyclical fluctuations often abstract from long-run economic growth. In particular, various business cycle models have been devised in which output fluctuates around a constant level of output rather than a path that grows over time. This approach again reflects the view that long-run growth is driven by forces that are independent of the factors that drive booms and busts in economic activity. On this assumption, we can analyse why output deviates from its long-run trend without bothering to model the trend itself.

While this dichotomy has proven useful for exploring certain questions, economists have become increasingly critical of this approach. Various attempts at integrating these two phenomena can be found in work on growth and business cycles from the late 1960s and early 1970s. Richard Goodwin's entry for growth and cycles in the first edition of this dictionary surveys some of this work (Goodwin 1987, pp. 574–6).

Arguably, however, the article that contributed most to advancing the view that growth and business cycles should be analysed within a single model is Kydland and Prescott (1982). They argued that business cycles were driven not by short-run variations in aggregate demand, as most previous work had assumed, but by fluctuations in the same force that drives long-run growth, namely, technological progress. They started with the Ramsey (1928) growth model in which long-run growth is due to labour-augmenting technical change. But rather than assuming a constant rate of technical change, they allowed it to vary over time. This captures the notion that new ideas arrive sporadically, so productivity growth is inherently random. Households react to these shocks by varying their capital accumulation and labour supply.

Kydland and Prescott went on to argue that technology shocks could account for most of the volatility in US output during the post-war period. This claim remains controversial. However, even those who were sceptical of the claim that productivity shocks were responsible for business cycles were forced to acknowledge that temporary shocks could affect decisions relevant for long-run growth, such as capital accumulation, and conversely that the forces which shape long-run growth could have important short-run consequences. This implies that treating growth and cycles as distinct processes

might overlook important connections between the two phenomena.

While Kydland and Prescott's paper was influential in promoting the view that growth and cycles should be modelled jointly, their model offered only limited insight into the connection between the two. This is because they modelled both long-run growth and fluctuations as exogenous: output per capita in their model grows because the economy is assumed to undergo technical change, and it grows in cycles because technical change is assumed to occur in cycles. As such, their model does not explain what drives technical change, why it should be inherently volatile, or whether growth and cycles might affect one another.

For example, Kydland and Prescott's model cannot tell us whether business cycles affect the rate of long-run growth. Are entrepreneurs more reluctant to undertake activities that lead to technical change in the face of macroeconomic volatility? Addressing this question requires us to model growth as an endogenous process rather than as the outcome of exogenous technical change. As another example, Kydland and Prescott asserted that technical change is inherently volatile. While this is undoubtedly true for any individual sector, it is not obvious why this volatility does not cancel out in the aggregate, resulting in a constant rate of technical change for the economy as a whole. Addressing this question requires us to model the underlying fluctuations in the rate of technical change as an endogenous outcome rather than as the result of an exogenous process. Fortunately, economists have since developed models in which either long-run growth or fluctuations, or both, are endogenous.

One line of research endogenizes growth while maintaining exogenous fluctuations. This approach allows us to study the effects of cyclical fluctuations on long-run growth. One of the first papers to tackle this question was Leland (1974), who built on previous work by Levhari and Srinivasan (1969). The latter studied the problem of a household deciding between consumption and saving given uncertain returns on its savings. Leland showed that this model could be reinterpreted as a representative household economy with a linear technology for producing output from capital. Growth in this model was driven by capital accumulation, so shocks to productivity – the analogue of uncertain returns – affected growth by affecting average investment.

Leland showed that the effect of cycles on growth depended on household attitudes towards risk. If the coefficient of relative risk aversion among households exceeded one, they would engage in more precautionary savings in the face of macroeconomic volatility, accumulating capital more rapidly. When relative risk aversion is below unity, macroeconomic volatility would induce households to accumulate less capital, leading to a slower rate of growth. Thus, whether cycles encourage or discourage long-run growth is ambiguous from a theoretical standpoint.

Ramey and Ramey (1995) provided empirical evidence on the relationship between growth and cycles using cross-country evidence. They found that volatility is associated with slower growth. At the same time, they found that more volatile countries do not have lower investment rates, contradicting Leland's analysis on how volatility ought to affect growth. This contradiction was resolved by Ramey and Ramey (1991) and Barlevy (2004), who argued that volatility affects growth not by changing average investment but by making investment less volatile; more volatile investment lowers long-run growth because growth is concave in investment. Barlevy (2004) in particular argued that this channel implies that exogenous cyclical fluctuations would be associated with very large welfare costs.

A separate line of research proceeded in the opposite direction: it assumed long-run growth was exogenous, and examined whether fluctuations in the economy-wide rate of technical change could arise endogenously. For example, Shleifer (1986) developed a multi-sector model where in each period innovators in a fixed fraction of sectors develop more productive technologies. They could use these to earn profits for a limited period, after which rivals in their sector could copy the technology and drive profits to zero. If innovators implemented their ideas as soon as they came up with them, the rate of aggregate technical change would be constant. But Shleifer allowed firms to delay implementation, and showed that there exist

equilibria where technical change occurs in spurts: innovators wait until there is enough innovation in other sectors before they implement their own ideas, so growth would be concentrated rather than spread out evenly over time.

Shleifer's result emerges because in his model implementing new technologies exhibits strategic complementarities: when one firm implements a new technology, its owners earn excess profits which they use to purchase goods in other sectors. Firms that come up with a new technology might therefore prefer to wait until others come up with new ideas. Even though the economy arrives at new ideas at a constant rate, growth proceeds at an uneven rate in equilibrium.

A third line of research has sought to endogenize both long-run growth and fluctuations. For example, Francois and Lloyd-Ellis (2003) consider a modification of the Shleifer model where innovators choose how much research to undertake, rather than assuming the rate at which new ideas arrive is fixed exogenously. This allows them to examine whether implementation cycles can affect long-run growth. Since implementation cycles emerge endogenously, the connection between cycles and growth may be different from the way growth responds to exogenous shocks as in Leland's analysis.

Francois and Lloyd-Ellis find that the equilibrium with cycles involves unambiguously higher average growth than the equilibrium in which innovators implement their new ideas immediately. This is because innovators earn higher profits when they coordinate implementation, providing them with more incentive to engage in research. However, welfare turns out to be lower in the presence of cycles, so faster but more uneven growth is less desirable. Lastly, Francois and Lloyd-Ellis show that, if countries differ in research productivity, we would observe a negative correlation between growth and cycles across countries; countries that are less productive in research will grow more slowly and exhibit longer and larger deviations from average growth. This helps to reconcile their results with Ramey and Ramey's evidence, and points out an important caveat for interpreting the cross-country evidence on growth and cycles.

Other authors have used models where both growth and cycles arise endogenously to explore whether technical change occurs in spurts not because of implementation delays but because of fluctuations in innovation. That is, even if innovators implement their new ideas immediately, they might still choose to concentrate their research activity in particular periods. Examples include Bental and Peled (1996), Walde (2002), and Matsuyama (1999). All three describe models in which the economy alternates between capital accumulation and innovation. In the first two papers, successful innovation raises the marginal product of capital, inducing a shift towards capital accumulation until the return to capital is low enough for innovation to turn profitable again. Matsuyama develops a model in which the economy grows as the variety of goods produced increases. Profits depend on the ratio of capital to the number of goods, so successful innovation reduces the profitability of innovation rather than increase the returns to physical capital. But all three models imply that the amount of innovation, and thus the rate of technical change, fluctuates along the equilibrium path.

A central feature of these models is that the economy fluctuates between innovation and capital accumulation. However, empirical evidence suggests research and development activity is high when capital accumulation is high. Recent work by Comin and Gertler (2006) and Barlevy (2005) examines why research activity might vary positively with capital accumulation. However, both assume cycles are due to exogenous shocks rather than that they arise endogenously in equilibrium. It remains a question for future research whether innovation might fluctuate endogenously but still co-vary with capital accumulation.

See Also

- ▶ Endogenous Growth Theory
- ▶ Kydland, Finn Erling (1943–)
- ▶ Prescott, Edward Christian (Born 1940)
- ▶ Real Business Cycles
- ▶ Technical Change
- ▶ Welfare Costs of Business Cycles

Bibliography

Barlevy, G. 2004. The cost of business cycles under endogenous growth. *American Economic Review* 94: 964–990.

Barlevy, G. 2005. *On the timing of innovation in stochastic schumpeterian growth models*. Working paper. Chicago: Federal Reserve Bank of Chicago.

Bental, B., and D. Peled. 1996. The accumulation of wealth and the cyclical generation of new technologies: A search theoretic approach. *International Economic Review* 37: 687–718.

Comin, D., and M. Gertler. 2006. Medium term business cycles. *American Economic Review* 96 (3): 523–551.

Francois, P., and H. Lloyd-Ellis. 2003. Animal spirits through creative destruction. *American Economic Review* 93: 530–550.

Goodwin, R. 1987. Growth and cycles. In *The new Palgrave: A dictionary of economics*, vol. 2, ed. J. Eatwell, M. Milgate, and P. Newman. Basingstoke: Palgrave, 1998.

Kydland, F., and E. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50: 1345–1370.

Leland, H. 1974. Optimal growth in a stochastic environment. *Review of Economic Studies* 41: 75–86.

Levhari, D., and T. Srinivasan. 1969. Optimal savings under uncertainty. *Review of Economic Studies* 36: 153–163.

Matsuyama, K. 1999. Growing through cycles. *Econometrica* 67: 1617–1631.

Ramey, G., and V. Ramey 1991. *Technology commitment and the cost of economic fluctuations*. Working paper no. 3755. Cambridge, MA: NBER.

Ramey, G., and V. Ramey. 1995. Cross-country evidence on the link between volatility and growth. *American Economic Review* 85: 1138–1151.

Ramsey, F. 1928. A mathematical theory of saving. *Economic Journal* 38: 543–559.

Shleifer, A. 1986. Implementation cycles. *Journal of Political Economy* 94: 1163–1190.

Walde, K. 2002. The economic determinants of technology shocks in a real business cycle model. *Journal of Economic Dynamics and Control* 27: 1–28.

Growth and Inequality (Macro Perspectives)

Vincenzo Quadrini

Abstract

This article provides a review of studies that examine the relationship between economic growth and inequality. These studies are divided into two groups. The first emphasizes the channels through which inequality affects growth, while the second emphasizes the opposite channel, where economic growth affects inequality. Although several empirical studies find a significant correlation between inequality and growth, it is still an open question as to whether the correlation is driven by the first or the second channel.

Keywords

Constitutions, economic approach to; Creative destruction; Distortionary taxes; Economic growth; Education; Endogenous growth theory; Entrepreneurship; Equity–efficiency trade-off; Expropriation; Growth and inequality; Human capital; Income smoothing; Increasing returns; Inequality; Innovation; Kuznets, S.; Policy reform; Redistribution of income and wealth; Skill-biased technical change; Wage inequality

JEL Classifications

D4; D10

The study of 'growth and inequality' has a long tradition. One well-known relationship in economic development is the Kuznets curve. Kuznets observed that in the early stage of human development, when agriculture was the main economic activity, inequality in the distribution of income was relatively low. As the economy industrialized and the workforce moved towards industry and away from agriculture, the distribution of income tended to widen. At some critical point in the economy's development, this tendency reversed. Although more recent evidence does not support the Kuznets curve hypothesis (see, for example, the widening income inequality observed in the United States starting in the early 1980s), the original empirical finding of Kuznets (1955) stimulated a large body of research activity.

Other evidence of the relationship between inequality and growth comes from more recent cross-country observations. Data collected during the 1980s and 1990s show that there is a great deal of variation across countries in the degree of income inequality and economic growth. Are the

different growth rates related to the degree of inequality of each individual country? Several studies find, indeed, that there is a cross-country negative correlation between inequality and growth, that is, countries with greater income inequality tend to experience slower growth; see Benabou (1996) and Perotti (1996) for a review of the empirical studies. But correlation does not imply causation, and there are good reasons to think that the causation can go in both directions. In other words, slow growth could generate greater inequality and equality could lead to faster growth.

Inequality Affecting Growth

One of the channels through which inequality affects growth is through the political and institutional system. A new series of studies in the 1980s, pioneered by Romer (1986) and Lucas (1988), developed a new class of models in which government policies could have a significant impact on the long-term growth of the economy (endogenous growth models). Given the importance of government policies for long-term growth, it becomes important to understand the forces and mechanisms underlying the choice of policies. This work stimulated a new series of studies in political economy. These studies start from the observation that, in a democratic society, the fundamental mechanism underlying the choice of policies is the electoral system. Therefore, in order to understand how policies are selected, we need to study the policy preferences of the population and how these preferences are translated into voting preferences.

Many factors affect the voting preferences of a society. However, for policies that have a clear redistributive content, the position of the voter in the distribution of income or wealth plays an important role. If a person is poor, his or her tax payments are smaller than the benefits he or she receives from government expenditures. Consequently, his attitude towards redistributive policies is more favourable than someone at the top of the income distribution (he or she has to pay more taxes than the received benefits). If the distribution of income is very unequal, then there will be many voters favouring larger governments. Of course this would not be a problem for efficiency if taxes were not distortionary. But, in a standard endogenous growth model, taxes have a negative impact on investment and growth. Therefore, the main conclusion of this literature is that inequality impairs the economic potential of a country because voters will demand more redistribution through distortionary taxes; Persson and Tabellini (1994), Alesina and Rodrik (1994), Krusell and Rìos-Rull (1996), Krusell et al. (1996) are some examples.

These studies also demonstrate the importance of the institutional system. Although greater inequality implies a greater demand for redistributive policies, the way political preferences are aggregated and the way policies are ultimately chosen depend on the particular institutional framework. For example, whether the representative democracy works through a parliamentary or a presidential system could lead to different sizes of government and, through distortionary taxes, to different levels of economic growth; see Persson and Tabellini (2005) for an analysis of the economic effects of constitutions.

The predictions of the politico-economic literature are consistent with several empirical studies as they find a negative relation between inequality and growth. However, a deeper empirical investigation of this channel poses some doubts. More specifically, the politico-economic channel can be divided into two sub-channels: a positive relation between ‘inequality’ and ‘redistributive policies’ and a negative relation between ‘redistributive policies’ and ‘growth’. Perotti (1996) shows that the negative effect of redistributive policies on growth is not a robust feature of the data. On the contrary, redistributive policies may even be positively associated with economic growth. How is this possible?

Several theories envision a beneficial effect of redistributive taxes. The key ingredient is the presence of financial constraints. Let us take the Shumpeterian view that entrepreneurship is central to economic growth. However, due to financial constraints and the lack of insurance markets, entrepreneurial investment is suboptimal. Under

these conditions, redistribution may provide extra resources to constrained entrepreneurs and could facilitate more investments in growth-enhancing activities. At the same time, a redistributive system provides an implicit system of income smoothing (a person pays high taxes when he or she earns high profits but receives payments in case of losses), and therefore, it provides insurance. If entrepreneurs are risk averse, this encourages more investment. The issue of whether redistributive taxes increase or decrease entrepreneurial investment is still an open area of research.

A similar story applies to investment in education or human capital. If education is important for economic growth, but because of financial constraints households choose sub-optimal levels of education, then government transfers may allow for greater investment and growth. A more direct effect could be generated by financing public education, as in Glomm and Ravikumar (1992). Examples of studies that emphasize the importance of inequality for growth in the presence of financial constraints are Galor and Zeira (1993), Banerjee and Newman (1993) and Aghion and Bolton (1997).

Another group of studies emphasizes social conflict and expropriation. Greater inequality means that a larger group of individuals is at the bottom of the distribution and faces poor economic conditions compared to the rest of the population. Faced with poor economic conditions, people have strong incentives to expropriate either by 'stealing' or through 'revolutions'. The risk of expropriation has two negative effects. First, it acts as an investment tax that discourages investment. Second, more resources are devoted to protect property rights, which detracts from resources devoted to productive and growth enhancing activities. An example of this kind of theory is Benhabib and Rustichini (1996).

Another theory of inequality affecting growth is that developed in Murphy et al. (1989). This theory assumes that there are technologies with increasing returns. These technologies become profitable only if the domestic market is sufficiently large, that is, there is a large demand for the goods produced with the new technologies. If wealth is highly concentrated the domestic market remains small (since there are not enough consumers who can afford these goods). As a result, these growth-enhancing technologies will not be implemented. However, the theory finds weak support in the data (see Benabou 1996).

Growth Affecting Inequality

If we take the view that growth requires innovative risky activities and these activities cannot be easily insured, we would expect that faster growth is associated with greater *ex post* inequality. At the same time, a faster rate of innovation implies greater destructions of monopoly positions (creative destruction). This would generate lower inequality because the monopoly positions, which are the source of high-income revenues, last for a shorter period of time. Therefore, it is not obvious whether faster innovation and growth create greater inequality. However, within this environment, faster growth generates higher mobility due to a higher turnover in the holding of monopoly positions. Therefore, even if growth leads to greater inequality, it also creates a healthier social environment. Long-term growth requires technological innovation and there is no doubt that new technologies affect different groups in different ways. Therefore, growth and inequality are intrinsically related. Since 1980, wage inequality among different education groups has been widening in almost all industrialized countries. Katz and Murphy (1992) show that this increase is due to a raising demand of skilled labour. Krusell et al. (2000) propose an explanation for the increasing demand of skilled labour based on the introduction and development of new technologies that are more complementary to skilled labour (skill-biased technologies).

Suppose that there are two types of workers, skilled and unskilled. The stocks of skilled and unskilled workers change slowly over time. Now suppose that there is the introduction of skill-biased technologies, that is, technologies that require more skilled labour than unskilled labour. This will lead to an increase in the demand for skilled workers. Given the limited increase in the supply, the wages of skilled workers will increase.

On the other hand, the demand for unskilled workers will decline, which leads to a fall in the wages of these workers.

This is a compelling explanation for the increasing wage premium started at the beginning of the 1980s. However, it raises the question of why the ratio of skilled versus unskilled workers has not increased that much during this period, certainly not as much as we would expect given the size of the wage premium change.

The technological innovations introduced in the 1970s seem to have affected the economy in other respects. Greenwood and Jovanovic (1999) and Hobijn and Jovanovic (2001) believe that new information technologies required a level of restructuring that incumbent firms could not face. As a result, their stock market value dropped. This is another form of redistribution in the sense that the owners of incumbent firms lose market value to the owners of the new firms.

Policy Considerations

Whether we concentrate on the first channel of causation – in which growth affects inequality – or to the second – in which inequality affects growth – there are no obvious policy recommendations. If we think that inequality has a negative impact on growth because society demands more redistributive policies (as in the standard political economy literature), then the constitutional system of electoral representation becomes central. Changing the constitutional system could lead to different political outcomes. However, changing the constitutional system is not easy. We could also think of reallocating resources once and for all to change the initial distribution. Although this is possible in theory, it is difficult from a political point of view.

If we concentrate on the opposite channel, in which growth impacts on inequality, and we are concerned about having an excessively unequal distribution of incomes, then we may consider possible redistributive policies. However, these policies may also have undesired effects on efficiency. If the tax system keeps the after-tax skill premium low, the incentive to acquire skills will be lower. But, because of skilled-biased technologies, more skills are required. This could also discourage the introduction of these technologies, which would impact negatively on growth. The equity–efficiency trade-off becomes central to the analysis. The positioning of a society in this trade-off will depend on society preferences about the degree of inequality that is socially acceptable. These preferences are based on individual beliefs that are likely to depend on individual experiences and they change very slowly over time. The relationship between personal experience and beliefs is formalized in Piketty (1995); see also Quadrini (1999).

See Also

- ▶ Economic Growth
- ▶ Inequality (International Evidence)
- ▶ Policy Reform, Political Economy of
- ▶ Wage Inequality, Changes in

Bibliography

Aghion, P., and P. Bolton. 1997. A theory of trickle-down growth and development. *Review of Economic Studies* 64: 151–172.

Alesina, A., and D. Rodrik. 1994. Distributive politics and economic growth. *Quarterly Journal of Economics* 109: 465–490.

Banerjee, A.V., and A.F. Newman. 1993. Occupational choice and the process of development. *Journal of Political Economy* 101: 274–298.

Benabou, R. 1996. Inequality and growth. In *NBER macroeconomics annual*, ed. B.S. Bernanke and J. Rotemberg. Cambridge, MA: MIT Press.

Benhabib, J., and A. Rustichini. 1996. Social conflict and growth. *Journal of Economic Growth* 1: 129–146.

Galor, O., and J. Zeira. 1993. Income distribution and macroeconomics. *Review of Economic Studies* 60: 35–52.

Glomm, G., and B. Ravikumar. 1992. Public versus private investment inhuman capital: Endogenous growth and income inequality. *Journal of Political Economy* 100: 818–834.

Greenwood, J., and B. Jovanovic. 1999. The information technology revolution and the stock market. *American Economic Review* 89: 116–122.

Hobijn, B., and B. Jovanovic. 2001. The IT revolution and the stock market: Evidence. *American Economic Review* 91: 1203–1220.

Katz, L., and K. Murphy. 1992. Changes in relative wages, 1963–1987: Supply and demand factors. *Quarterly Journal of Economics* 107: 35–78.
Krusell, P., and J.V. Rìos-Rull. 1996. Vested interests in a positive theory of stagnation and growth. *Review of Economic Studies* 63: 301–331.
Krusell, P., V. Quadrini, and J.-V. Rìos-Rull. 1996. Are consumption taxes really better than income taxes? *Journal of Monetary Economics* 37: 475–503.
Krusell, P., L. Ohanian, J.-V. Rìos-Rull, and G. Violante. 2000. Capital–skill complementarity and inequality: A macro economic analysis. *Econometrica* 68: 1029–1053.
Kuznets, S. 1955. Economic growth and income inequality. *American Economic Review* 45: 1–28.
Lucas, R.E. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22: 3–42.
Murphy, K., A. Shleifer, and R. Vishny. 1989. Income distribution, market size and industrialization. *Quarterly Journal of Economics* 104: 537–564.
Perotti, R. 1996. Growth, income distribution, and democracy: What the data say. *Journal of Economic Growth* 1: 149–187.
Persson, T., and G. Tabellini. 1994. Is inequality harmful for growth? *American Economic Review* 84: 600–621.
Persson, T., and G. Tabellini. 2005. *The economic effects of constitutions*. Cambridge, MA: MIT Press.
Piketty, T. 1995. Social mobility and redistributive politics. *Quarterly Journal of Economics* 110: 551–584.
Quadrini, V. 1999. Growth, learning and redistributive policies. *Journal of Public Economics* 74: 263–297.
Romer, P.M. 1986. Increasing returns and long-run growth. *Journal of Political Economy* 94: 1002–1036.

Growth and Institutions

Daron Acemoglu

Abstract

Institutions are often viewed as a key determinant of economic growth. Much research inquires whether the institutions that influence economic outcomes are themselves determined by other factors. European colonization of the world provides a laboratory in which to investigate these issues since it exogenously imposed different institutions on otherwise identical societies. Colonies where Europeans settled had institutions that protected property rights, and have since prospered, while other colonies were given centralized repressive states that extracted resources from the population and have largely remained relatively poor. Choice of institutions reflects the distribution of political power in a society.

Keywords

Colonization; Commitment; Growth and governance; Growth and institutions; Mortality; North, D.; Political institutions, economic approaches to; Property rights

JEL Classifications

O4

A central question of economics is to understand why some countries are much poorer than others. Economists have long recognized that this relates to the fact that some countries have much less human capital, physical capital and technology than others, and use their existing factors and technologies much less efficiently. Nevertheless, these differences are only proximate causes in the sense that they pose the next question of why some countries have less human capital, physical capital and technology and make worse use of their factors and opportunities. This has motivated economists and social scientists more broadly to look for potential fundamental causes, which may be underlying these proximate differences across countries.

Institutions have emerged as a potential fundamental cause, contrasting, for example, with geographical differences or cultural factors. While geographic characteristics of countries and regions may lead to differences in the technology available to individuals or make their investments in physical and human capital more difficult, institutional differences, associated with differences in the organization of society, shape economic and political incentives and affect the nature of equilibria via these channels. There is vibrant research, both empirical and theoretical, attempting to understand the importance of institutions for economic outcomes. Since it is impossible to do justice to this burgeoning field in such a short article, my purpose here is not to survey the

literature but to present some of the main conceptual issues that are useful for future work.

What Are Institutions?

Douglass North (1990, p. 3) offers the following definition: 'Institutions are the rules of the game in a society or, more formally, are the humanly devised constraints that shape human interaction.' Three important features of institutions are apparent in this definition: (a) they are 'humanly devised', which contrasts with other potential fundamental causes like geographic factors, which are outside human control; (b) they are 'the rules of the game' setting 'constraints' on human behavior; and (c) their major effect will be through incentives (see also North 1981).

The notion that incentives matter is second nature to economists, and institutions, if they are a key determinant of incentives, should have a major effect on economic outcomes, including economic development, growth, inequality and poverty. But do they? Are institutions key determinants of economic outcomes or secondary arrangements that respond to other, perhaps geographic or cultural, determinants of human and economic interactions?

Much empirical research attempts to answer this question. Before we discuss some of this research, it is useful to emphasize an important point: ultimately, the aim of the research on institutions is to pinpoint specific institutional characteristics that are responsible for economic outcomes in specific situations (for example, the effect of legal institutions on the types of business contracts). However, the starting point is often the impact of a broader notion of institutions on a variety of economic outcomes. This broader notion, in line with Douglass North's conception, incorporates many aspects of the economic, political and social organization of society. Institutions can differ between societies because of their formal methods of collective decision-making (democracy versus dictatorship) or because of their economic institutions (security of property rights, entry barriers, the set of contracts available to businessmen). They may also differ because a given set of formal institutions is expected to, and does, function differently; for example, they may differ between two societies that are democratic because the distribution of political power lies with different groups or social classes, or because in one society democracy is expected to collapse while in the other it is consolidated. This broad definition of institutions is both an advantage and a curse. It is an advantage, since it enables us to get started with theoretical and empirical investigations of the role of institutions without getting bogged down by taxonomies. It is a curse, since, unless we can follow it up with a better understanding of the role of specific institutions, we have learned little.

The Impact of Institutions

There are tremendous cross-country differences in the way that economic and political life is organized. A voluminous literature documents large cross-country differences in economic institutions, and a strong correlation between these institutions and economic performance. Knack and Keefer (1995), for instance, look at measures of property rights enforcement compiled by international business organizations, Mauro (1995) looks at measures of corruption, Djankov et al. (2002) compile measures of entry barriers across countries, while many studies look at variation in educational institutions and the corresponding differences in human capital. All of these authors find substantial differences in these measures of economic institutions, and significant correlation between these measures and various indicators of economic performance. For example, Djankov et al. find that, while the total cost of opening a medium-size business in the United States is less than 0.02% of GDP per capita in 1999, the same cost is 2.7% of GDP per capita in Nigeria, 1.16% in Kenya, 0.91% in Ecuador and 4.95% in the Dominican Republic. These entry barriers are highly correlated with various economic outcomes, including the rate of economic growth and the level of development.

Nevertheless, this type of correlation does not establish that the countries with worse institutions

are poor because of their institutions. After all, the United States differs from Nigeria, Kenya and the Dominican Republic in its social, geographic, cultural and economic fundamentals, so these may be the source of their poor economic performance. In fact, these differences may be the source of institutional differences themselves. Consequently, evidence based on correlation does not establish whether institutions are important determinants of economic outcomes.

To make further progress, one needs to isolate a source of exogenous differences in institutions, so that we approximate a situation in which a number of otherwise identical societies end up with different sets of institutions. European colonization of the rest of the world provides a potential laboratory in which to investigate these issues. From the late 15th century, Europeans dominated and colonized much of the rest of the globe. Together with European dominance came the imposition of very different institutions and social power structures in different parts of the world.

Acemoglu et al. (2001) document that in a large number of colonies, especially those in Africa, Central America, the Caribbean and South Asia, European powers set up 'extractive states'. These institutions (again broadly construed) did not introduce much protection for private property, nor did they provide checks and balances against the government. The explicit aim of the Europeans in these colonies was extraction of resources, in one form or another. This colonization strategy and the associated institutions contrast with the institutions Europeans set up in other colonies, especially in colonies where they settled in large numbers: for example, the United States, Canada, Australia and New Zealand. In these colonies the emphasis was on the enforcement of property rights for a broad cross section of the society, especially smallholders, merchants and entrepreneurs. The term 'broad cross section' is emphasized here since, even in the societies with the worst institutions, the property rights of the elite are often secure, but the vast majority of the population enjoys no such rights and faces significant barriers to participation in many economic activities. Although investments by the elite can generate economic growth for limited periods, for sustained growth property rights for a broad cross section seem to be crucial (Acemoglu et al. 2002; Acemoglu 2003).

A crucial determinant of whether Europeans chose the path of extractive institutions was whether they settled in large numbers. In colonies where Europeans settled, the institutions were developed for their own future benefits. In colonies where Europeans did not settle, their objective was to set up a highly centralized state apparatus, and other associated institutions, to oppress the native population and facilitate the extraction of resources in the short run. Based on this idea, Acemoglu et al. (2001) suggest that, in places where the disease environments made it easy for Europeans to settle, the path of institutional development should have been different from areas where Europeans faced high mortality rates.

In practice, during the time of colonization, Europeans faced widely different mortality rates in colonies because of differences in the prevalence of malaria and yellow fever. These mortality rates provide a possible candidate for a source of exogenous variation in institutions. The mortality rates should not directly influence output today but, by affecting the settlement patterns of Europeans, they may have had a first-order effect on institutional development. Consequently, these potential settler mortality rates can be used as an instrument for broad institutional differences across countries in an instrumental-variables estimation strategy.

The key requirement for an instrument is that it should have no direct effect on the outcome that is the object of interest (other than its effect via the endogenous regressor). There are a number of channels through which potential settler mortality could influence current economic outcomes or may be correlated with other factors influencing these outcomes. Nevertheless, there are also good reasons why, as a first approximation, these mortality rates should not have a direct effect. Malaria and yellow fever were fatal to Europeans who had no immunity, and thus had a major effect on settlement patterns, but they had much more limited effects on natives who, over centuries, had

developed various types of immunities. The exclusion restriction is also supported by the death rates of native populations, which appear to be similar between areas with very different mortality rates for Europeans (see, for example, Curtin 1964).

The data also show that there were major differences in the institutional development of the high-mortality and low-mortality colonies. Moreover, consistent with the key idea in Acemoglu et al. (2001), various measures of broad institutions – for example, measures of protection against expropriation – are highly correlated with the death rates Europeans faced more than a century ago and with early European settlement patterns. They also show that these institutional differences induced by mortality rates and European settlement patterns have a major (and robust) effect on income per capita. For example, the estimates imply that improving Nigeria's institutions to the level of those in Chile could, in the long run, lead to as much as a sevenfold increase in Nigeria's income. This evidence suggests that, once we focus on potentially exogenous sources of variation, the data point to a large effect of broad institutional differences on economic development.

Naturally, mortality rates faced by Europeans were not the only determinant of Europeans' colonization strategies. Acemoglu et al. (2002) focus on another important aspect, namely, how densely different regions were settled before colonization. They document that in more densely settled areas Europeans were more likely to introduce extractive institutions because it was more profitable for them to exploit the indigenous population, either by having them work in plantations and mines or by maintaining the existing system and collecting taxes and tributes. This suggests another source of variation in institutions that may have persisted to the present, and Acemoglu et al. (2002) show similar large effects from this source of variation.

Another example that illustrates the consequences of difference in institutions is the contrast between North Korea and South Korea. The geopolitical balance between the Soviet Union and the United States following the Second World War led to separation along the 38th parallel. The North, under the dictatorship of Kim II Sung, adopted a very centralized command economy with little role for private property. In the meantime, South Korea, though far from a free-market economy, relied on a capitalist organization of the economy, with private ownership of the means of production and legal protection for a range of producers, especially those under the umbrella of the chaebols, the large family conglomerates that dominated the South Korean economy. Although not democratic during its early phases, the South Korean state was generally supportive of rapid development and is often credited with facilitating, or even encouraging, investment and rapid growth in Korea.

Under these two highly contrasting regimes, the economies of North and South Korea diverged. While South Korea grew rapidly under capitalist institutions and policies, North Korea has experienced minimal growth since 1950 under communist institutions and policies.

Overall, a variety of evidence paints a picture in which broad institutional differences across countries have had a major influence on their economic development. This evidence suggests that to understand why some countries are poor we should understand why their institutions are dysfunctional. But this is only part of a first step in the journey towards an answer. The next question is even harder: if institutions have such a large effect on economic riches, why do some societies choose, end up with and maintain these dysfunctional institutions?

Modelling Institutional Differences

As a first step in modelling institutions, let us consider the relationship between three institutional characteristics: (a) economic institutions, (b) political power, and (c) political institutions.

As already mentioned, economic institutions matter for economic growth because they shape the incentives of key economic actors in society; in particular, they influence investments in physical and human capital and technology, and the organization of production. Economic institutions determine not only the aggregate economic

growth potential of the economy but also the distribution of resources in the society, and herein lies part of the problem: different institutions will be associated not only with different degrees of efficiency and potential for economic growth, but also with different distributions of the gains across different individuals and social groups.

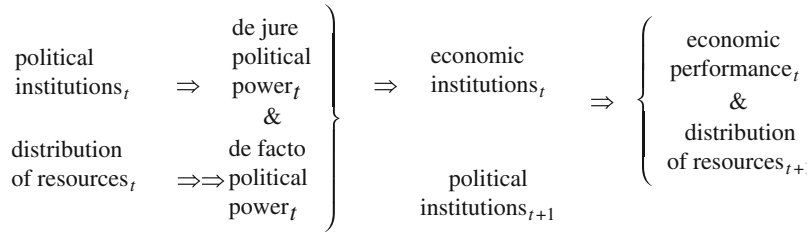
How are economic institutions determined? Although various factors play a role here, including history and chance, ultimately economic institutions are produced by collective choices of the society. And because of their influence on the distribution of economic gains, not all individuals and groups typically prefer the same set of economic institutions. This leads to a conflict of interest among various groups and individuals over the choice of economic institutions; and the political power of the different groups will be the deciding factor.

The distribution of political power in society is also endogenous. To make more progress here, let us distinguish between two components of political power; de jure (formal) and de facto political power (see Acemoglu and Robinson 2006). De jure political power refers to power that originates from the political institutions in society. Political institutions, similar to economic institutions, determine the constraints on and the incentives of the key actors, but this time in the political sphere. Examples of political institutions include the form of government – for example, democracy versus dictatorship or autocracy – and the extent of constraints on politicians and political elites.

A group of individuals, even if they are not allocated power by political institutions, may possess political power; for example, they can revolt, use arms, hire mercenaries, co-opt the military, or undertake protests in order to impose their wishes on society. This type of de facto political power originates from both the ability of the group in question to solve its collective action problem and from the economic resources available to the group (which determine their capacity to use force against other groups).

This discussion highlights the fact that we can think of political institutions and the distribution of economic resources in society as two state variables, affecting how political power will be distributed and how economic institutions will be chosen. An important notion is that of persistence; the distribution of resources and political institutions are relatively slow-changing and persistent. Since, like economic institutions, political institutions are collective choices, the distribution of political power in society is the key determinant of their evolution. This creates a central mechanism of persistence: political institutions allocate de jure political power, and those who hold political power influence the evolution of political institutions, and they will generally opt to maintain the political institutions that give them political power. A second mechanism of persistence comes from the distribution of resources: when a particular group is rich relative to others, this will increase its de facto political power and enable it to push for economic and political institutions favorable to its interests, reproducing the initial disparity. Despite these tendencies for persistence, the framework also emphasizes the potential for change. In particular, 'shocks' to the balance of de facto political power, including changes in technologies and the international environment, have the potential to generate major changes in political institutions, and consequently in economic institutions and economic growth.

Acemoglu et al. (2005b) summarize this framework in Fig. 1.



Growth and Institutions, Fig. 1

Institutions in Action

As a brief example, consider the development of property rights in Europe during the Middle Ages. Lack of property rights for landowners, merchants and protoindustrialists was detrimental to economic growth during this epoch. Since political institutions at the time placed political power in the hands of kings and various types of hereditary monarchies, such rights were largely decided by these monarchs. The monarchs often used their powers to expropriate producers, impose arbitrary taxation, renege on their debts, and allocate the productive resources of society to their allies in return for economic benefits or political support. Consequently, economic institutions during the Middle Ages provided little incentive to invest in land, physical or human capital, or technology, and failed to foster economic growth. These economic institutions also ensured that the monarchs controlled a large fraction of the economic resources in society, solidifying their political power and ensuring the continuation of the political regime.

The 17th century, however, witnessed major changes in the economic and political institutions that paved the way for the development of property rights and limits on monarchs' power, especially in England after the civil war of 1642–6 and the Glorious Revolution of 1688, and in the Netherlands after the Dutch revolt against the Hapsburgs. How did these major institutional changes take place? In England until the 16th century the king also possessed a substantial amount of de facto political power, and, if we leave aside civil wars related to royal succession, no other social group could amass sufficient de facto political power to challenge the king. But changes in the English land market (Tawney 1941) and the expansion of Atlantic trade in the 16th and 17th centuries (Acemoglu et al. 2005a) gradually increased the economic fortunes, and consequently the de facto power, of landowners and merchants opposed to the absolutist tendencies of the Kings.

By the 17th century, the growing prosperity of the merchants and the gentry, based on both internal and overseas (especially Atlantic) trade, enabled them to field military forces capable of defeating the king. This de facto power overcame the Stuart monarchs in the English civil war and Glorious Revolution, and led to a change in political institutions that stripped the king of much of his previous power over policy. These changes in the distribution of political power led to major changes in economic institutions, strengthening the property rights of both landowners and capital owners and spurring a process of financial and commercial expansion. The consequence was rapid economic growth, culminating in the industrial revolution,and a very different distribution of economic resources from that in the Middle Ages.

This discussion poses, and also gives clues about the answers to, two crucial questions. First, why do the groups with conflicting interests not agree on the set of economic institutions that maximize aggregate growth? Second, why do groups with political power want to change political institutions in their favour? In the context of the example above, why did the gentry and merchants use their de facto political power to change political institutions rather than simply implement the policies they wanted? The issue of commitment is at the root of the answers to both questions.

An agreement on the efficient set of institutions is often not forthcoming because of the complementarity between economic and political institutions and because groups with political power cannot commit to not using their power to change the distribution of resources in their favour. For example, economic institutions that increased the security of property rights for landowners and capital owners during the Middle Ages would not have been credible as long as the monarch monopolized political power. He could promise to respect property rights, but then at some point renege on his promise, as exemplified by the numerous financial defaults by medieval kings. Credible secure property rights necessitated a reduction in the political power of the monarch. Although these more secure property rights would foster economic growth, they were not appealing to the monarchs, who would thereby lose their rents from predation and expropriation as well as

various other privileges associated with their monopoly of political power. This is why the institutional changes in England as a result of the Glorious Revolution were not simply conceded by the Stuart kings. James II had to be deposed for the changes to take place.

The reason why political power is often used to change political institutions is related. In a dynamic world, individuals care about not only economic outcomes today but also those in the future. In the example above, the gentry and merchants were interested in their profits and therefore in the security of their property rights, not only in the present but also in the future. Therefore, they would have liked to use their (de facto) political power to secure benefits in the future as well as the present. However, commitment to future allocations (or economic institutions) is in general not possible because decisions in the future are made by those who hold political power at the time. If the gentry and merchants had been certain to maintain their de facto political power, this would not have been a problem. However, de facto political power is often transient, for example because the collective action problems that are solved to amass this power are likely to resurface in the future, or other groups, especially those controlling de jure power, can become stronger in the future. Therefore, any change in policies and economic institutions that relies purely on de facto political power is likely to be reversed in the future. In addition, many revolutions are followed by conflict among the revolutionaries. Recognizing this, the English gentry and merchants strove not just to change economic institutions in their favour following their victories against the Stuart monarchy, but also to alter political institutions and the future allocation of de jure power. Using political power to change political institutions then emerges as a useful strategy to make gains more durable. Consequently, political institutions and changes in political institutions are important as ways of manipulating future political power, and thus indirectly shaping future, as well as present, economic institutions and outcomes. Acemoglu and Robinson (2000, 2006) and Acemoglu et al. (2005b) provide more detailed models and discuss further applications, including the creation and consolidation of electoral democracies in the West and in Latin America.

See Also

- ▶ Class
- ▶ Growth and Civil War
- ▶ Political Institutions, Economic Approaches to

Bibliography

Acemoglu, D. 2003. *The form of property rights: Oligarchic versus democratic societies*, Working paper, vol. 10037. Cambridge, MA: NBER.

Acemoglu, D., S. Johnson, and J. Robinson. 2001. The colonial origins of comparative development: An empirical investigation. *American Economic Review* 91: 1369–1401.

Acemoglu, D., S. Johnson, and J. Robinson. 2002. Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics* 118: 1231–1294.

Acemoglu, D., S. Johnson, and J. Robinson. 2005a. The rise of Europe: Atlantic trade, institutional change and economic growth. *American Economic Review* 95: 546–579.

Acemoglu, D., S. Johnson, and J. Robinson. 2005b. Institutions as the fundamental cause of long-run growth. In *Handbook of economic growth*, ed. P. Aghion and S. Durlauf. Amsterdam: North-Holland.

Acemoglu, D., and J. Robinson. 2000. Why did the West extend the franchise? Democracy, inequality and growth in historical perspective. *Quarterly Journal of Economics* 115: 1167–1199.

Acemoglu, D., and J. Robinson. 2006. *Economic origins of dictatorship and democracy*. Cambridge: Cambridge University Press.

Curtin, P. 1964. *The image of Africa*. Madison: University of Wisconsin Press.

Djankov, S., R. LaPorta, F. Lopez-de-Silanes, and A. Shleifer. 2002. The regulation of entry. *Quarterly Journal of Economics* 117: 1–37.

Knack, S., and P. Keefer. 1995. Institutions and economic performance: Cross-country tests using alternative measures. *Economics and Politics* 7: 207–227.

Mauro, P. 1995. Corruption and growth. *Quarterly Journal of Economics* 110: 681–712.

North, D. 1981. *Structure and change in economic history*. New York: W.W. Norton and Co.

North, D. 1990. *Institutions, institutional change, and economic performance*. New York: Cambridge University Press.
Tawney, R. 1941. The rise of the gentry, 1558–1640. *Economic History Review* 11: 1–38.

Growth and International Trade

James Rauch

Abstract

International trade is believed to promote growth for countries at the technological frontier by expanding the market over which to exploit new ideas. For countries behind the technological frontier, three views of the relationship between growth and international trade are described and assessed empirically: trade hampers growth for natural resource-abundant countries that specialize in the export of technologically stagnant primary products; trade acts as the 'handmaiden' of growth by improving the quality of investment and slowing the tendency of its return to fall; and trade acts as the engine of growth by providing a conduit for technology transfer.

Keywords

Comparative advantage; Economies and diseconomies of scale; Foreign direct investment (FDI); Global commodity chains; Growth and international trade; Heckscher-Ohlin-Samuelson model; International trade and technology; Learning by doing; Less developed countries; Primary product exports; Research and development; Ricardo-Viner model; Rybczynski effect; Size of economies; Technological progress; Technology transfer

JEL Classifications

O4

When examining the interaction between international trade and economic growth, it is necessary to distinguish between countries that are at the technological frontier and those that are substantially behind it. The reason is that long-run growth of productivity and ultimately of per capita income is constrained by technological progress in the former group of countries but not in the latter group.

Growth and International Trade for Countries at the Technological Frontier

Rivera-Batiz and Romer (1991) argue that a larger economic size of the world promotes worldwide technological progress, in two ways. First, current research and development (R&D) builds upon the existing stock of ideas or knowledge (standing on the shoulders of giants). Because ideas are nonrival, a larger world implies a larger stock of knowledge that facilitates R&D. International trade is implicated in this effect only to the extent that it promotes exchange or sharing of ideas among countries. Second, a larger economic size of the world raises the rents generated by monopoly holders of patented ideas (or ideas that are too costly to imitate) by providing a larger market for the goods based on the ideas. International trade is needed to exploit this large market and therefore provides greater incentives for innovation, increasing economic growth through this 'Schumpeterian' mechanism. Rivera-Batiz and Romer concentrate on these scale effects by modelling integration between similar, developed countries, thereby abstracting from the comparative advantage, resource reallocation effects of trade.

This theoretical prediction of a positive effect of international trade on the worldwide rate of technological progress needs to be qualified when international trade occurs between dissimilar countries. For example, trade can be expected to increase the relative price of skilled labour in the most skilled labour-abundant country. This could raise the cost of R&D relative to the cost

of goods production enough to offset the positive effects of international trade on R&D for this country. If the most skilled labour-abundant country drives the worldwide rate of technological progress, international trade could reduce world economic growth. For a much fuller discussion of the interactions between international trade, R&D, and economic growth, see Grossman and Helpman (1991).

In so far as technological progress is transmitted equally to all countries at the worldwide technological frontier, there is no room for cross-country variation in the impact of international trade on growth from this source. In other words, the prediction that trade increases the worldwide rate of technological progress (absent strong, offsetting comparative advantage effects), and hence long-run economic growth of countries at the technological frontier, is inherently untestable using crosssectional data because we have only one world. Since falling transportation and communication costs have tended to make countries more open to international trade over time, this prediction is consistent with time-series evidence that seems to show that the worldwide rate of growth is increasing over the very long run, but this increase is also predicted by the increasing size of the world – standing on the shoulders of more giants.

Growth and International Trade for Countries Behind the Technological Frontier

The effect of international trade on the economic growth of less developed countries (LDCs) has long been one of the most passionately debated subjects in economics. Here I set out three views that span the range from negative to positive.

Trade as the Enemy of Growth

The case for trade as the enemy of growth can be made succinctly using the open economy version of the model of Matsuyama (1992). In his model all productivity increase takes place through learning by doing in the manufacturing sector, which is perceived as an external economy by any individual manufacturing firm. (In Matsuyama's model, unlike in the classic infant-industry argument, the potential for learning by doing is never exhausted – the manufacturing sector never catches up to international best practice. One way this can be justified is by assuming that international best practice is a receding target). Productivity in agriculture (or, more broadly, in the primary product sector) is constant by assumption. Now consider a country that is well endowed with natural resources relative to labour (and also relative to human and physical capital, if these are assumed to be used more intensively in manufacturing). Comparative advantage will lead this country to export primary products and import manufactures. Exports of primary products draw workers out of manufacturing, both directly and indirectly by generating rents that create demand for services, and thereby reduce productivity growth both in manufacturing and in the aggregate (since there is no productivity growth in primary products). (A variant of this argument is that trade causes 'lagging' economies to specialize in goods whose learning potential has been exhausted: Young 1991). In this way trade reduces growth in per capita income in countries with abundant natural resources. It is interesting that the sociological literature on 'dependency' and 'world systems' comes to the same conclusion that development of 'peripheral' countries is hindered by their exports of primary products to 'core' industrialized countries. For a summary of these arguments and a review of empirical studies see Crowly et al. (1998).

This case for trade as the enemy of growth does not depend on the assumption that productivity in the primary product sector is constant. Obviously learning-by-doing and other forms of productivity increase occur in this sector in the real world. What is crucial is only that this productivity increase tends to be substantially less rapid or less sustainable for a long period than in manufacturing. It is also possible that productivity growth in the primary product exportable sector worsens the terms of trade for the exporting country rather than raising its income, as

modelled by Lewis (1969). In this instance the country in question must be large enough in its primary product speciality to influence its world price, or else we must treat the countries to which the argument applies as a bloc rather than individually.

Recently it has become fashionable to introduce inequality and 'institutions' into the argument linking trade to slow growth through a comparative advantage in primary products. In this view the problem with exports of primary products, especially minerals and tropical cash crops, is not that they are associated with low productivity growth but rather that they are associated with high inequality between owners of large mines or plantations and their employees. This inequality leads to the adoption of educational and political institutions that tend to exclude and disenfranchise the masses, making these economies less capable of realizing the potential offered by new technologies. This argument has been most thoroughly articulated by Engerman and Sokoloff (2002).

Several studies (for example, Sachs and Warner 2001) find that the ratio of primary product exports to GDP is strongly negatively associated with per capita income growth. This negative association remains even after many other potential determinants of economic growth are controlled for, such as climate, geography, economic policies, political institutions, and external shocks. Indeed, the ratio of primary product exports to GDP is considered one of the most robust determinants of growth in cross-country growth regressions (Sala-i-Martin 1997). This cross-country evidence cannot be considered decisive, however. Easterly (2001) points out that the primary product export share for less developed countries has tended to decline since 1960 as resources have been depleted and population has grown, yet per capita income growth rates have not tended to rise.

Trade as the 'Handmaiden' of Growth

The phrase 'trade as handmaiden of growth' is taken from the title of an article by Kravis (1970). He writes (1970, p. 869),

> The term 'engine of growth' is not generally descriptive and involves expectations which cannot be fulfilled by trade alone; the term 'handmaiden of growth' better conveys the notion of the role that trade can play. One of the most important parts of this handmaiden role for today's developing countries may be to serve as a check on the appropriateness of new industries by keeping the price and cost structures in touch with external prices and costs.

This supportive role can be usefully compared to that of financial development. As discussed in Levine (1997), a well-developed financial system increases the efficiency of investment by helping to channel savings to the most profitable projects. One way that trade can increase the efficiency of investment is by helping to ensure that the most privately profitable projects are also the most socially profitable ones. Foreign competition discourages investors from attempting to establish monopoly positions in small domestic markets and from producing substandard goods. Other ways in which trade can increase the efficiency of investment are enabling producers to realize economies of scale through exporting, and relieving bottlenecks that might reduce the returns to well-conceived downstream investments or divert resources from them. (Openness to international trade can also generate static and dynamic economies of scale – the latter through learning-by-doing spillovers, for example – by promoting specialization. Weinhold and Rauch (1999), find that productivity growth in the manufacturing sector in less developed countries is higher when production is more specialized).

Along the same lines, trade can slow or suspend the tendency of the return on investment to fall as physical (or human) capital accumulates. This link between growth and international trade was originally made by Ricardo, who argued that repeal of the Corn Laws in Britain would increase imports of grain, reduce the competition for labour between agricultural landlords and manufacturers, and thereby raise the return to investment in reproducible physical capital. His implicit model was what we now call the Ricardo–Viner model. Here we consider an extension of this model by Deardorff (1984), in which a small open economy produces an agricultural good using land and labour and,

potentially, a number of manufactured goods using (reproducible) capital and labour under conditions of constant returns to scale and perfect competition. Let the manufactured goods be ranked unambiguously by capital intensity. Now consider a less developed country with an endowment of capital relative to labour and land that is so small that its manufacturing sector is completely specialized in production of the least capital-intensive manufactured good. As the country accumulates capital its return (marginal product) will fall. However, this reduction in the return to capital allows the country to become internationally competitive in the next least capital-intensive manufactured good, so that its manufacturing sector becomes incompletely specialized. At this point the return to capital (and the wage and, by extension, the rent on land) becomes fixed by international goods prices, as in the standard Heckscher–Ohlin–Samuelson model. Further accumulation of capital then causes both capital and labour resources within the manufacturing sector to be reallocated from the less to the more capital-intensive good (the Rybczynski effect). This forestalls any fall in the return to capital until the manufacturing sector becomes completely specialized in the more capital-intensive good, after which the return to capital falls until production of the next most capital-intensive good is introduced, and so on.

If the view that international trade is the handmaiden of economic growth is correct, the large empirical literature investigating 'causality' between trade and growth (for example, Jung and Marshall 1985) is somewhat beside the point. There should, however, be a strong cross-country correlation between openness to international trade and the rate of growth. Many studies have found such a correlation, but its robustness has been called into question (Rodriguez and Rodrik 2001). Part of the problem may be that most studies try to include all countries for which there are reliable data, yet we have seen that there is no clearly predicted relationship between openness to trade and economic growth for countries at the technological frontier (the so-called industrialized or rich countries) and that trade may reduce growth for countries whose exports are concentrated in primary products. The positive correlation between openness to trade and growth should be most robust for the intermediate group of countries between least and most developed, a group sometimes labelled the 'semi-industrialized' countries. Given the complexity of the handmaiden view, however, it is probably best investigated by theoretically informed case studies like the classic NBER volumes (1974–8) supervised by Jagdish N. Bhagwati and Anne O. Krueger.

Trade as the Engine of Growth

The view of trade as the engine of growth takes technological progress rather than investment to be the ultimate source of growth, and sees imported ideas as the main determinant of technological progress in less developed countries. In other words, trade with more technologically advanced countries acts as a vehicle for the flow of knowledge from them and thereby drives growth in less advanced countries. Foreign direct investment (FDI) from more to less developed countries plays the same role. (This contemporary view of trade as the engine of growth must be distinguished from the older view, in which growth is driven by expansion of land devoted to production of technologically stagnant primary products to meet the demand of industrialized countries; see, for example, Caves 1965). The emphasis on imported ideas is associated with the work of Romer (for example, 1993). His work, however, leaves open the question of the specific mechanisms through which firms in less developed countries absorb knowledge from contact with technologically advanced countries.

Economists have typically modelled technology transfer as an arm's-length phenomenon. Firms are not *taught* the new technology. Rather, they engage in purposive imitative activity on their own (see, for example, Grossman and Helpman 1991, ch. 11), employ machinery and equipment that embodies foreign knowledge (for example, Coe et al. 1997), license the new technology, and so on. In reality, however, it is difficult to learn new technology from a distance. Keller (2004, p. 756) writes, 'Only the broad outlines of technological knowledge are codified – the remainder remains "tacit". non-codified

knowledge is usually transferred through person-to-person demonstrations and instructions.' There is a growing body of evidence that, for less developed country firms in particular, a major and perhaps predominant source of technology transfer (and transfer of managerial know-how) is instruction by developed country buyers: producers seeking cheaper suppliers of inputs and distributors seeking cheaper suppliers of final goods.

One example of such evidence is a study by Egan and Mody, who surveyed US buyers operating in LDCs, including 'manufacturers, retailers, importers, buyers' agents, and joint venture partners' (1992, p. 322). They found:

> Buyers also render long-term benefits to suppliers in the form of information on production technology. This occurs principally through various forms of inplant training. The buyer may send international experts to train local workers and supervisors ... Buyers may also arrange short-term worker training in a developed country plant. (1992, p. 328)

Rhee, Ross-Larson and Pursell surveyed Korean exporters of manufactures. Their findings were similar to those of Egan and Mody:

> The relations between Korean firms and the foreign buyers went far beyond the negotiation and fulfillment of contracts. Almost half the firms said they had directly benefited from the technical information foreign buyers provided: through visits to their plants by engineers or other technical staff of the foreign buyers, through visits by their engineering staff to the foreign buyers ... (1984, p. 61)

This process of learning foreign technology can be thought of as taking place within international production networks or 'global commodity chains' (Gereffi 1994, 1999). This theoretical framework predicts that, once LDC firms are incorporated into the 'bottoms' of the chains, their learning will continue by movement up the chains. There are two types of chains: 'producer-driven' and 'buyer-driven' (Gereffi 1994, p. 97). In the former, large manufacturers play the central roles in coordinating the production networks. Producer-driven chains are typical in capital- and technology-intensive industries such as automobiles, aircraft, computers, semiconductors, and heavy machinery. In the latter, large retailers, branded marketers, and branded manufacturers play the coordinating roles. Buyer- driven commodity chains are typical in labour-intensive, consumer goods industries such as garments, footwear, toys, housewares, and consumer electronics. Profitability is highest at the tops of the chains where barriers to entry are greatest: scale and technology in producer-driven chains, design and marketing expertise in buyer- driven chains.

In buyer-driven commodity chains, one mode through which learning is predicted to continue is *organizational succession*: from assembler to original equipment manufacturer (OEM) to original brand-name manufacturer (OBM), which is from more subordinate, competitive, and low-profit positions to more controlling, oligopolistic, high-profit positions. In the apparel industry, Gereffi (1999) finds that LDC firms that have parts provided to them for assembly learn how to find on their own the parts needed to make the product according to the design specified by the buyer (and may then subcontract the assembly); firms that have reached this level learn how to design and sell their own merchandise, becoming branded manufacturers (and may then subcontract the production, becoming branded marketers). Additional study is needed to determine whether this pattern of learning is common in other consumer goods industries. At the same time, work is needed to reconcile the kind of findings discussed here with econometric analyses (surveyed in. Rodrik 1999, ch. 2) which conclude that more productive firms export, but exporting does not make firms more productive.

In producer-driven commodity chains, one mode of learning is through 'vertical linkages' established between foreign subsidiaries of the large manufacturers that coordinate the production networks and host country suppliers. Saggi (2002, p. 213) writes:

> Mexico's experience with FDI is illustrative of how such a process works. In Mexico, extensive backward linkages resulted from FDI in the automobile industry. Within five years of investments by major auto manufacturers there were 300 domestic producers of parts and accessories, of which 110 had annual sales of more than $1 million (Moran 1998). Foreign producers also transferred industry best practices, zero defect procedures, and production audits to domestic suppliers, thereby improving their productivity and the quality of their products.

Javorcik (2004) finds econometric evidence that in Lithuania upstream suppliers to foreign subsidiaries experienced increases in productivity.

Conclusions

Here I highlight what I feel are the most important challenges facing the exponents of each of the three views of trade and growth described in the previous section. Those who see trade as the enemy of growth for natural resource-abundant countries need to do more than merely assert that the primary product export sector is incapable of rapid productivity growth. This assumption has recently been challenged by Dolan et al. (1999) and others who provide evidence that the fresh vegetable export sector in sub-Saharan Africa can realize the kind of learning benefits and investment opportunities associated with manufacturing by upgrading quality and presentation. Those who see trade as the handmaiden of growth must formulate their views more precisely if they are to be used to guide policy or subjected to rigorous empirical testing. Finally, those who see trade as the engine of growth must resolve the contradictions between case studies and econometric results regarding the benefits from exporting. Are the case studies unrepresentative, or does the statistical estimation suffer from measurement problems (as suggested by Katayama et al. 2003)? Perhaps the case studies and the surveys that collect the data for econometric analysis need to be coordinated to make sure that the right questions are being asked.

See Also

- ▶ Growth and Learning-By-Doing
- ▶ International Trade Theory
- ▶ Schumpeterian Growth and Growth Policy Design

Bibliography

Caves, R. 1965. 'Vent for surplus' models of trade and growth. In *Trade, growth, and the balance of payments*, ed. R. Baldwin et al. Chicago: Rand McNally.

Coe, D., E. Helpman, and A. Hoffmaister. 1997. North–South R&D spillovers. *Economic Journal* 107 (440): 134–149.

Crowly, A., J. Rauch, S. Seagrave, and D. Smith. 1998. Quantitative cross-national studies of economic development: A comparison of the economics and sociology literatures. *Studies in Comparative International Development* 33: 30–57.

Deardorff, A. 1984. An exposition and exploration of Krueger's trade model. *Canadian Journal of Economics* 17: 731–746.

Dolan, C., C. Harris-Pascal, and J. Humphrey 1999. *Horticulture commodity chains: The impact on the UK market of the African fresh vegetable industry*. Working Paper No. 96. Brighton: Institute of Development Studies.

Easterly, W. 2001. The lost decades: Developing countries' stagnation in spite of policy reform 1980–1998. *Journal of Economic Growth* 6: 135–157.

Egan, M.-L., and A. Mody. 1992. Buyer-seller links in export development. *World Development* 20: 321–334.

Engerman, S., and K. Sokoloff. 2002. Factor endowments, inequality, and paths of development among new world economies. *Economia* 3: 41–88.

Gereffi, G. 1994. The organization of buyer-driven global commodity chains: How U.S. retailers shape overseas production networks. In *Commodity chains and global capitalism*, ed. G. Gereffi and M. Korzeniewicz. Westport: Praeger.

Gereffi, G. 1999. International trade and industrial upgrading in the apparel commodity chain. *Journal of International Economics* 48: 37–70.

Grossman, G., and E. Helpman. 1991. *Innovation and growth in the global economy*. Cambridge, MA: MIT Press.

Javorcik, B. 2004. Does foreign direct investment increase the productivity of domestic firms? In search of spillovers through backward linkages. *American Economic Review* 94: 605–627.

Jung, W., and P. Marshall. 1985. Exports, growth, and causality in developing countries. *Journal of Development Economics* 18: 1–12.

Katayama, H., S. Lu, and J. Tybout 2003. *Why plant-level productivity studies are often misleading, and an alternative approach to inference*. Working Paper No. 9617. Cambridge, MA: National Bureau of Economic Research.

Keller, W. 2004. International technology diffusion. *Journal of Economic Literature* 42: 752–782.

Kravis, I. 1970. Trade as a handmaiden of growth: Similarities between the nineteenth and twentieth centuries. *Economic Journal* 80: 850–872.

Levine, R. 1997. Financial development and economic growth: Views and agenda. *Journal of Economic Literature* 35: 688–726.

Lewis, W. 1969. *Aspects of tropical trade, 1883–1965*. Stockholm: Almqvist and Wicksell.

Matsuyama, K. 1992. Agricultural productivity, comparative advantage, and economic growth. *Journal of Economic Theory* 58: 317–334.

Moran, T. 1998. *Foreign direct investment and development*. Washington, DC: Institute for International Economics.
NBER (National Bureau of Economic Research). 1974–78. *Foreign trade regimes and economic development*. New York: Columbia University Press.
Rhee, Y., B. Ross-Larson, and G. Pursell. 1984. *Korea's competitive edge: Managing the entry into world markets*. Baltimore: Johns Hopkins University Press.
Rivera-Batiz, L., and P. Romer. 1991. Economic integration and endogenous growth. *Quarterly Journal of Economics* 106: 531–555.
Rodriguez, F., and D. Rodrik. 2001. Trade policy and economic growth: A skeptic's guide to the cross-national evidence. In *NBER macroeconomics annual 2000*, ed. B. Bernanke and K. Rogoff. Cambridge, MA: MIT Press.
Rodrik, D. 1999. *Making openness work: The new global economy and the developing countries*. Washington, DC: Overseas Development Council.
Romer, P. 1993. Idea gaps and object gaps in economic development. *Journal of Monetary Economics* 32: 543–573.
Sachs, J., and A. Warner. 2001. The curse of natural resources. *European Economic Review* 45: 827–838.
Saggi, K. 2002. Trade, foreign direct investment, and international technology transfer: A survey. *World Bank Research Observer* 17: 191–235.
Sala-i-Martin, X. 1997. I just ran two million regressions. *American Economic Review* 87: 178–183.
Weinhold, D., and J. Rauch. 1999. Openness, specialization, and productivity growth in less developed countries. *Canadian Journal of Economics* 32: 1009–1027.
Young, A. 1991. Learning by doing and the dynamic effects of international trade. *Quarterly Journal of Economics* 106: 369–405.

Growth and Learning-By-Doing

Paul Beaudry

Keywords
Arrow, K.; Economic growth theory; Learning by doing; Lucas, R.; Productive efficiency

JEL Classifications
O4

Learning by doing refers to improvements in productive efficiency arising from the generation of experience obtained by producing a good or service. The formal modelling of learning by doing was initiated in Arrow (1962) and was motivated by two main factors. The first motivating factor was empirical: several studies of wartime production found that input requirements decreased as a result of production experience. For example, Searle (1945) studied productivity changes in the Second World War shipbuilding programmes. During the Second World War, US production of ships increased dramatically, from 26 vessels in 1939 to 1,900 ships in 1943, an almost fiftyfold increase. Searle (1945) noticed that unit labour requirements decreased at a constant rate for a given percentage increase in output. On average, a doubling of output was associated with declines of 16 to 22 per cent in the number of man-hours required to build Liberty ships, Victory ships, tankers and standard cargo vessels. Alchian (1963) studied the relationship between the amount of direct labour required to produce an airframe and the number of airframes produced in the United States during the Second World War. He found that a doubling of production experience decreased labour input by approximately one-third. Other empirical studies of learning by doing include Rapping (1965), Irwin and Klenow (1994) and Thornton and Thompson (2001).

The second motivating factor behind the work of Arrow (1962) was a search for a theory of economic growth which did not rely on exogenous change in productivity as a driving force. In particular, Arrow's contribution and its extensions in Levhari (1966a, b) were to show how economic growth could be sustained in a market with perfect competition. Arrow's original model is quite sophisticated, but the main insight can be derived in a simpler setting, as shown in Sheshinski (1967) and presented here. Consider a one good economy, where the production of the good requires capital and labour input according to the constant returns to scale production function:

$$Y = F(K, AL), F(\lambda K, A\lambda L) = \lambda F(K, AL).$$

In this specification of the production technology, A represents the efficiency of labour in producing the good. The main idea in the learning by

doing literature is that A is a function of past experience. Arrow assumed that experience can be measured by cumulative investment or, in other words, the capital stock. The form of the relationship between A and the capital stock is posited to be:

$$A_t = (K_t)^{\alpha}, 0 < \alpha < 1$$

where the assumption that $0 < \alpha < 1$ is motivated by the empirical studies. In order to close the system, assume that the labour force grows exponentially at the rate η and let capital accumulation be driven by a constant saving rate out of incomes, s where, in the absence of depreciation, this implies

$$\dot{K} = sY$$

In this environment, on the assumption that the change in A is an unintended consequence of production, it can be shown that a balanced growth path exists where per-capita income and per-capita capital grow at the rate

$$\alpha \frac{\eta}{1-\alpha}$$

The two important aspects to note about the resulting growth rate is that it is positive if $\eta > 0$ and it is independent of the savings rate s. The additional property – that the rate of growth of income is tied to a positive rate of population growth – is generally seen as a weakness of this type of model. This property can be partially remedied, as shown in Romer (1986), if one assumes that $\alpha = 1$. In this case, even in the absence of labour force growth there exists a balanced growth path where the rate of growth is given by

$$sF(1, L)$$

The drawback of this specification ($\alpha = 1$) is that the growth rate now depends on the size of the labour force, which is referred to as a 'scale effect'. The attractive feature of this specification is that the growth rate can be modified by an economic decision variable such as the savings rate. An alternative way of modifying Arrow's original model is to posit, as in Lucas (1988), that A depends on the per-capita value of the capital stock instead of on the level of the capital stock. This assumption is justified in Lucas (1988) on the grounds that A reflects the knowledge of the average worker with respect to how best to operate the technology. In the case where the relationship is given by $A = \frac{K}{L}$, the steady growth rate of per-capita output is given by $sF(1, 1)\ \eta$. This formulation has the attractive property that it is positive even if $\eta = 0$, and it does not exhibit a scale effect. Accordingly it offers a succinct theory of economic growth. Lucas conjectured that the assumption of constant returns to learning (that is, $\alpha = 1$) could be justified in a model where there is bounded learning in any one good but where there is continual entry of new goods over time. This idea is formally studied in Stokey (1988) and Young (1993). There is also a large literature that discusses how learning by doing can interact with international trade and potentially give rise to income divergence across countries; see for example Lucas (1993) and Young (1991).

Bibliography

Alchian, A. 1963. Reliability of progress curves in airframe production. *Econometrica* 31: 679–693.

Arrow, K. 1962. The economic implications of learning by doing. *Review of Economic Studies* 29: 155–173.

Irwin, D., and P. Klenow. 1994. Learning by doing spillovers in the semiconductor industry. *Journal of Political Economy* 102: 1200–1227.

Levhari, D. 1966a. Further implications of learning by doing. *Review of Economic Studies* 33: 31–38.

Levhari, D. 1966b. Extensions of arrow's 'Learning by Doing'. *Review of Economic Studies* 33: 117–131.

Lucas, R. Jr. 1988. On the mechanics of economic development. *Journal of Monetary Economics* 22: 3–42.

Lucas, R. Jr. 1993. Making a miracle. *Econometrica* 61: 251–272.

Rapping, L. 1965. Learning and World War II production functions. *Review of Economics and Statistics* 47: 81–86.

Romer, P. 1986. Increasing returns and long run growth. *Journal of Political Economy* 94: 1002–1037.

Searle, A. 1945. Productivity of labour and industry. *Monthly Labor Review* 61: 1132–1147.

Sheshinski, E. 1967. Optimal accumulation with learning by doing. In *Essays on the theory of economic growth*, ed. K. Shell. Cambridge, MA: MIT Press.
Stokey, N. 1988. Learning by doing and the introduction of new goods. *Journal of Political Economy* 96: 701–717.
Thornton, R., and P. Thompson. 2001. Learning from experience and learning from others: An exploration of learning and spillovers in wartime shipbuilding. *American Economic Review* 91: 1350–1368.
Young, A. 1991. Learning by doing and the dynamic effects of international trade. *Quarterly Journal of Economics* 106: 369–405.
Young, A. 1993. Invention and bounded learning by doing. *Journal of Political Economy* 101: 443–472.

Growth Models, Multisector

W. A. Brock and W. D. Dechert

Abstract

Multisector growth models have been increasingly used since the 1980s. The duality between growth models and dynamic general equilibrium models renders the multisector growth model ideal for the analysis of efficient intertemporal resource allocation. This includes renewable and non-renewable natural resources, produced resources such as capital, and land and labour resources. Growth models have been widely used in business cycle theory and in asset pricing theory. They have also been applied to the optimal management of dynamic ecological systems that have an economic component as a part of a complex systems model.

Keywords

Asset pricing model; Bequest motive; Business cycles; Central limit theorem; Computation; Concavity; Convergence; Decentralization; Dynamic macroeconomic theory; Equity premium puzzle; Equivalence theorem; General equilibrium; Indirect utility function; Infinite horizons; Law of large numbers; Multisector growth models; New Keynesian macroeconomics; Optimal growth models; Optimal planning models; Overtaking ordering; Rational expectations equilibrium; Real business cycles; Recursive intertemporal general equilibrium models; Representative agent; Separating hyperplane theorem; Single-sector growth models; Turnpike theorems

JEL Classifications

O41

Multisector growth models are basic building blocks not only for optimal planning models (Majumdar 1987; McKenzie 1986) but also for recursive general equilibrium models (McKenzie 2002; Stokey and Lucas 1989), and for econometrically tractable models for business cycle research (Cooley 1995) and general macroeconomics (Sargent 1987). Majumdar (1987) has already covered some basic theory, some efficiency and decentralization analysis, as well as some optimization concepts. We attempt to fill in the space between Majumdar (1987) and the current research frontier as well as to outline applications not treated by Majumdar.

Before we begin, we wish to stress that the style of this article is to point the reader towards surveys of the subject in order to economize on references to the many researchers who have contributed to this rather large area, and to paint, in broad strokes, the overall structure of this research area, especially its impact on empirical work, in order to illuminate directions where the research frontier might go.

Dynamic macroeconomic theory has made much use of the stochastic one-sector growth model (Cooley 1995; Altug and Labadie 1994; Sargent 1987; Stokey and Lucas 1989), for two primary reasons. First, it is a classical result that optimal growth models can be viewed as general equilibrium models by use of the separating hyperplane theorem in an appropriate space to construct the support prices. See Becker and Boyd (1997) for this general result, which they call the 'equivalence theorem'. It is closely related to the use of decentralization prices in Majumdar (1987) and the general treatment of decentralization in Majumdar (1992).

The basic idea of the class of the 'equivalence theorem' of Becker and Boyd is as follows. Consider an infinite horizon intertemporal general equilibrium model with a representative infinitely lived consumer who faces intertemporal prices as given. Then it is a classical result that the rational expectations equilibrium of such a model is the same as the optimal solution of a planning problem where the planner has the same preferences as the representative consumer. Technical issues arise from the infinite horizon such as the necessity and sufficiency of transversality conditions at infinity (that is, the present discounted mathematical expectation of value of any stocks 'left over' at infinity should be zero, much as in a finite horizon case with no bequest motive). But the general ideas behind this type of result are much the same as in the well-known finite dimensional cases. See Becker and Boyd (1997) for the details.

Second, infinite horizon stochastic multisector models are also basic in constructing econometrically tractable models to use in analysing data. Here, especially, is where stochastic versions of the turnpike theorem (explained below) are used. For example, it is used to justify use of laws of large numbers and central limit theorems in econometric time-series applications.

A key property of the one-sector model that promotes its use in real business cycle applications as well as intertemporal general equilibrium asset pricing applications is the stochastic analog of the turnpike theorem. This theorem states that optimal capital stock and optimal consumption converge in a stochastic sense to a unique stochastic limit under standard assumptions of concavity of the payoff function (for example, the planner's preferences) and of the production function and modest assumptions on the structure of the stochastic shocks. It is much more difficult to obtain such results for general multisector stochastic models (Arkin and Evstigneev 1987; Marimon 1989) and even for deterministic versions of those models (McKenzie 1986, 2002).

However, one can show that if the discount rate on the future is small enough there are results available in the literature that locate useful sufficient conditions on payoffs and technology such that stochastic convergence occurs (Marimon 1989) and deterministic convergence occurs (McKenzie 1986, 2002). Results for stochastic multisector growth models in both discrete time settings and continuous time settings are also contained in the papers in Dechert (2001).

The basic idea behind these results, called 'turnpike' results, is to first observe that, if the discount rate on the future is zero, the dynamic optimization problem will attempt to maximize a long-run 'static' objective in order to avoid infinite 'value loss' if it failed to do so. Making this intuition mathematically precise requires introduction of a partial ordering called the 'overtaking ordering' and making assumptions on the objective function and the dynamics so that avoidance of infinite value loss results in convergence of the optimal quantities to a unique long-run limit (see Arkin and Evstigneev 1987, and the papers in Dechert 2001, for stochastic cases and McKenzie 1986, 2002, for deterministic cases.)

Once one has results well in hand for the case of zero discounting on the future, intuition suggests that there should be a notion of 'continuity' that would enable one to prove that, if the discount rate is close enough to zero, convergence would still hold. Unfortunately, turning such intuition into precise mathematics turns out to be rather difficult (see McKenzie 1986, 2002, for deterministic literature and Arkin and Evstigneev 1987, the papers in Dechert 2001, and Marimon 1989, for the stochastic case).

We attempt to give the reader a brief idea of how the mathematical arguments work in a sketch of the arguments used to prove turnpike theorems for the deterministic case below. Let preferences of a planner be given by

$$\max_{\{x_t\}} \sum_{t=0}^{\infty} \beta^t \left[u(x_t, x_{t-1}) - u\left(x_\beta^*, x_\beta^*\right) \right] \qquad (1)$$

where u: $R^{2n} \to R$ is a twice continuously differentiable function (typically an indirect utility or payoff function), β is a discount factor, $0 < \beta \leq 1$, and x_β^* is an optimal steady state which

solves the first-order necessary conditions of the optimization in Eq. (1):

$$D_1u(x_t, x_{t-1}) + \beta D_2u(x_{t+1}, x_t) = 0t \geq 1 \quad (2)$$

D_i denotes partial derivative with respect to the ith argument of u, and x_0 is given. We assume that u is jointly concave in its arguments and use Eq. (2) evaluated at the optimal steady state x^*_β to rewrite the sum in Eq. (1). To simplify the notation, we let $u^*_\beta = u\left(x^*_\beta, x^*_\beta\right)$ and $D_iu^*_\beta = D_iu\left(x^*_\beta, x^*_\beta\right)$. Also, define

$$d_t = -\left[u(x_t, x_{t-1}) - u^*_\beta - \left(D_1u^*_\beta\right)\left(x_t - x^*_\beta\right) - \left(D_2x^*_\beta\right)\left(x_{t-1} - x^*_\beta\right)\right]$$

which is positive by the concavity of u. With this notation,

$$\sum_{t=1}^{T} \beta^t\left[u(x_t, x_{t-1}) - u^*_\beta\right] = \beta^T\left(D_1u^*_\beta\right)\left(x_T - x^*_\beta\right) + \left(D_2x^*_\beta\right)\left(x_0 - x^*_\beta\right) - \sum_{t=1}^{T}\beta^t d_t \quad (3)$$

Equation (3) immediately suggests that a good strategy to construct candidate optimal programs $\{x_t\}$ is to choose a program $\{x_t\}$ to solve

$$\min \sum_{t=1}^{\infty} \beta^t d_t. \quad (4)$$

This strategy works for all $\beta \in (0, 1]$. Following McKenzie (1986, and his references to David Gale) for $\beta = 1$, classify a program $\{x_t\}$ as good (bad) if the series $\sum_{t=1}^{T} d_t$ converges (diverges) and note that all programs $\{x_t\}$ are either good or bad. Solve Eq. (4) over good programs to get a top candidate for an optimum. By defining an appropriate partial ordering of programs that is a total ordering on the set of good programs, this top candidate turns out to be optimum. Since the series $\{d_t\}$ converges to 0 for all good programs, this forces $\{x_t\}$ to converge to a unique x^* which is the maximizer of $u(x, x)$ under the assumption that u is strictly concave. We call this analytical strategy the 'value loss' strategy.

There are basically two analytical strategies used for the case β is less than but close to 1. It is beyond the scope of this article to discuss them here; see McKenzie (1986, 2002) for the details.

All three of these analytical strategies can be generalized to stochastic cases where the indirect utility u contains stochastic shocks provided that Markovian type conditions are assumed on the stock process; $\{x_t\}$ is replaced by a sequence of random variables $\{X_t\}$; and x^*_β is replaced by a certain stationary ergodic stochastic process, X^*_β, that plays the role of the optimal stochastic steady state. This is not simple but we hope that our outline of one of the analytical strategies makes that one, at least, intuitively plausible (see, for example, Arkin and Evstigneev 1987; Marimon 1989; and the papers by Brock and Majumdar 1978, Brock and Mirman 1972, and Brock and Magill 1979, reprinted in Dechert 2001).

Our sketch of the above results has been deliberately brief since excellent survey treatments are readily available in the literature that we have cited. We wish to discuss here applications of multisector models to the following areas of economics: (*a*) a general vision of how the economy works; (*b*) asset pricing; (*c*) coupled ecological/economic dynamical systems.

General Vision

It is no exaggeration to say that classical general equilibrium theory is analytically organized around existence of equilibrium, the core and equilibria, the two welfare theorems, as well as the 'anything goes' theorem of Sonnenschein, Mantel and Debreu (SMD) as the subject is expounded in McKenzie (2002). The SMD result requires users to place restrictions on the consumers and producers that populate general equilibrium models in order to use the theory for empirical work. In intertemporal economics a most popular way of doing this is to restrict

oneself to recursive intertemporal general equilibrium models, and that restriction (via the 'equivalence theorem') places us in the domain of multisector growth models (Becker and Boyd 1997).

Black (1995), stimulated by general equilibrium theory, sketches with broad strokes a vision of the economy that is basically operating close enough to a complete set of markets so that the device of generating equilibria by maximizing a weighted sum of utilities can be applied (McKenzie 2002). Analytically, this device puts us in the domain of a large multisector model viewed as general equilibrium via a generalization of the 'equivalence theorem' in Becker and Boyd (1997). As McKenzie (2002) shows, turnpike theory could be extended to recursive intertemporal general equilibrium models with heterogeneous consumers provided markets are complete. Black (1995) proposes adding various elements to received intertemporal recursive general equilibrium models (that is, multisector growth models) not only to fill in gaps in the existing literature up to the mid-1990s but also to make the models match up better to data.

The book by Altug et al. (2003) might be viewed as an example of a realization of Black's vision. It shows the power of variations on uses of single-sector and multisector growth models as building blocks for closed- and openeconomy macro models. We give some specific examples below. The examples are chosen because current cutting-edge work is being done in these areas and because the subject is moving fast in the directions of these chosen areas.

Asset Pricing

Use of the 'equivalence theorem' rapidly lead to development of recursive econometrically tractable intertemporal general equilibrium asset pricing models based upon multisector stochastic optimal growth models (Becker and Boyd, 1997, and the papers in Dechert, 2001). The confrontation with data has not been all positive. Three main directions in which these models failed when confronted with data came to be known as the equity premium family of puzzles. But Weitzman (2004, p. 1) has shown that '. . .the subjective distribution of the future growth rate has its mean and variance calibrated to average past values. This paper shows that using the Bayesian posterior estimates of these parameters can go a very long way toward eliminating simultaneously all three puzzles.' A major point of Weitzman is that, once the uncertainty inherent in the fact that there is estimation uncertainty in key parameters that the agents living in the model must take into account in addition to the shocks inherent in the model, then the puzzles tend to vanish.

Akdeniz and Dechert (2007) show that a single-sector stochastic asset pricing model with production and with heterogeneous firms can go a long ways toward removing the puzzles without having to introduce Weitzman's Bayesian modification of the underlying basic model. Work like that of Akdeniz and Dechert is now possible due to advances in computational technology. Jog and Schaller (1994) have shown that a modification of the basic model for liquidity-constrained firms can account for patterns of mean reversion observed in returns data across size classes of firms.

Macroeconomics

We have already mentioned the real business cycle literature (Cooley 1995; Altug et al. 2003) as macroeconomic applications of multisector growth models and their decentralization analysis. A major recent development in macroeconomics is to replace the representative consumer agent and competitive firms in such models with a representative agent facing a set of differentiated products, each produced by a differentiated products monopolist who faces a stochastic process that gives it realizations of periods when it is allowed to change prices. This strategic modeling device allows one to add an analytically tractable

theory of price setting which can be grafted onto the existing analytical apparatus of recursive multisector models to produce a model where a unification of the 'real side' and the 'monetary side' of macroeconomics can take place. Various devices are used to produce a demand for money balances in the model that include real balance services in the indirect utility function and cash in advance constraints. This modeling strategy has produced a new generation of very fruitful 'New Keynesian' macro models which has allowed treatment of key issues of monetary policy as well as better fit to data especially data resulting from interactions between the real side and the monetary side of an economy. See Altug et al. (2003) and, especially, Woodford's treatise (2003) for this genre.

The real world has distortions such as taxes, inflation and other government activities such as production of public goods which require modifications of the basic structure of intertemporal recursive general equilibrium theory. Fortunately the analytical core can be quite readily modified to include these elements (Turnovsky 1995).

Much of the literature on multisector optimal growth theory assumes convex technology and concave payoff (that is, concave utility) so that the indirect utility $u(x_t, x_{t-1}, S_t)$ is jointly concave in (x_t, x_{t-1}) for each value of the stochastic shock S_t. We believe much activity in the future will involve generalizations to models of coupled ecological and economic dynamic systems where such concavity does not hold. Some analytical work in this area has already appeared (Becker and Boyd 1997; Majumdar 1992) and as computational technology progresses we expect to see more developments that use a combination of analytics and computation.

See Also

- ▶ Intertemporal Equilibrium and Efficiency
- ▶ Rational Expectations
- ▶ Stochastic Optimal Control

Bibliography

Akdeniz, L., and W. Dechert. 2007. The equity premium in Brock's asset pricing model. *Journal of Economic Dynamics and Control* 31: 2263–2292.

Altug, S., J. Chadha, and C. Nolan. 2003. *Dynamic macroeconomic analysis: Theory and policy in general equilibrium*. Cambridge: Cambridge University Press.

Altug, S., and P. Labadie. 1994. *Dynamic choice and asset markets*. New York: Academic Press.

Arkin, V., and I. Evstigneev. 1987. *Stochastic models of control and economic dynamics*. New York: Academic Press.

Becker, R., and R. Boyd. 1997. *Capital theory, equilibrium analysis and recursive utility*. Oxford: Blackwell.

Black, F. 1995. *Exploring general equilibrium*. Cambridge, MA: MIT Press.

Brock, W.A., and M.J.P. Magill. 1979. Dynamics under uncertainty. *Econometrica* 47: 843–868.

Brock, W.A., and M. Majumdar. 1978. Global asymptotic stability results for multisector models of optimal growth under uncertainty when future utilities are discounted. *Journal of Economic Theory* 18: 225–243.

Brock, W.A., and L. Mirman. 1972. Optimal economic growth and uncertainty: the discounted case. *Journal of Economic Theory* 4: 479–513.

Cooley, T., ed. 1995. *Frontiers of business cycle research*. Princeton: Princeton University Press.

Dechert, W., ed. 2001. *Growth theory, nonlinear dynamics, and economic modelling: scientific essays of William Allen Brock*. Cheltenham: Edward Elgar.

Jog, V., and H. Schaller. 1994. Finance constraints and asset pricing: evidence on mean reversion. *Journal of Empirical Finance* 1: 193–209.

Majumdar, M. 1987. Multisector growth models. In *The new palgrave: A dictionary of economics*, ed. J. Eatwell, M. Milgate, and P. Newman. London: Macmillan.

Majumdar, M., eds. 1992. *Decentralization in infinite horizon economies*. Boulder: Westview Press.

Marimon, R. 1989. Stochastic turnpike property and stationary equilibrium. *Journal of Economic Theory* 47: 282–306.

McKenzie, L. 1986. Optimal economic growth, turnpike theorems, and comparative dynamics. In *Handbook of mathematical economics*, ed. K. Arrow and M. Intriligator, vol. 3. Amsterdam: North-Holland.

McKenzie, L. 2002. *Classical general equilibrium theory*. Cambridge, MA: MIT Press.

Sargent, T. 1987. *Dynamic macroeconomic theory*. Cambridge, MA: Harvard University Press.

Stokey, N., and R. Lucas. 1989. *Recursive methods in economic dynamics*. Cambridge, MA: MIT Press.

Turnovsky, S. 1995. *Methods of macroeconomic dynamics*. Cambridge, MA: MIT Press.

Weitzman, M. 2004. The Bayesian equity premium. Working paper, Department of Economics, Harvard University.

Woodford, M. 2003. *Interest and prices*. Princeton: Princeton University Press.

Growth Take-Offs

Matthias Doepke

Abstract

Following a phase of near-constant living standards lasting from Stone Age until the onset of the Industrial Revolution, a large number of countries have experienced growth takeoffs, in which stagnation gives way to sustained economic growth. What causes some countries to enter a growth takeoff while others remain poor? We discuss three mechanisms that can trigger a growth takeoff in a country previously trapped in poverty: fertility decline, structural change, and accelerating technological progress.

Keywords

Child labour; Demographic transition; Economic growth; Fertility; Fixed factors; Growth takeoffs; Human capital; Income–population feedback; Industrial revolution; Land; Malthus's theory of population; Mortality; Nutrition and development; Population growth; Productivity growth; Skill-intensive technology; Stagnation; Structural change; Technological progress; Women's work

JEL Classification

D4; D10

Viewed on a historical timescale, economic growth in the world economy is characterized by a long phase of stagnation in living standards, followed in many, but not all, countries by a growth take-off, that is, a transition to steady and sustained economic growth.

Figure 1 illustrates the basic facts. Before 1800, GDP per capita was low and near-constant in all world regions, with little cross-country variation in income levels. The first country to experience a growth take-off was Britain with the start of the Industrial Revolution, closely followed by other west European countries and the 'Western Offshoots' such as the United States. More recently, a number of Asian and Latin American countries have undergone a transition to rapid economic growth as well. In much of Africa, however, income per capita continues to stagnate. What causes some countries to enter a growth takeoff while others remain poor?

Explaining Stagnation

Before one can account for a growth takeoff after a phase of stagnation, it is essential to understand why economies stagnated in the first place. The explanation suggested by one of the earliest writers on the subject, British economist Thomas Malthus in his *Essay on the Principle of Population* of 1798, is widely accepted to the present day. The Malthusian model relies on two key ingredients: an agricultural production function that uses the fixed factor of land, and an income–population feedback where the population growth rate is an increasing function of income per capita.

Consider an aggregate production function of the form

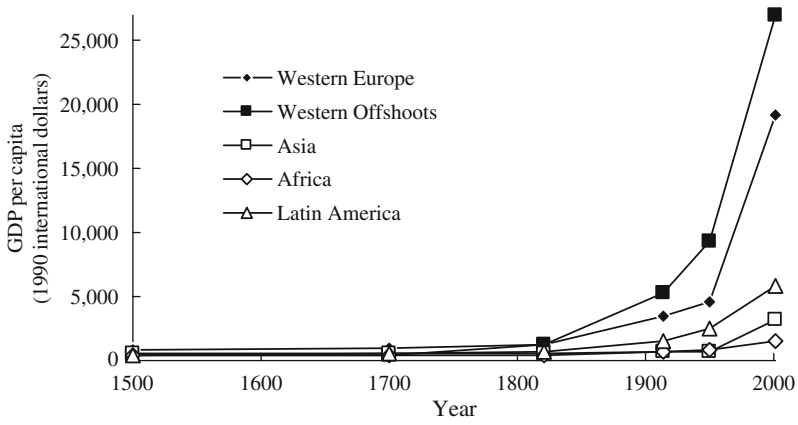
$$Y_t = A_t N_t^{\alpha} Z^{1-\alpha}, \tag{1}$$

where Y_t denotes output in period t, A_t is productivity, N_t is the size of the population, and Z is the fixed amount of land. (The results outlined below can be generalized to the case where physical capital also enters production.) In what follows, we use lower-case letters to denote per capita variables ($y_i = Y_t = N_t$, and so on), and the growth rate of a variable x is written as $\gamma(x)$. Output per capita is given by $y_t = Atz_t^{1-\alpha}$, so that its growth rate $\gamma(y_t)$ satisfies:

$$\gamma(y_t) = \gamma(A_t) + (1-\alpha)\gamma(z_t).$$

Since land Z is constant, we have $\gamma(z_t) = \gamma(Z) - \gamma(N_t) = -\gamma(N_t)$. Using this relationship, the growth equation can be rewritten as

$$\gamma(y_t) = \gamma(A_t) - (1-\alpha)\gamma(z_t). \tag{2}$$



Growth Take-Offs, Fig. 1 The evolution of income per capita across world regions, years 1500–2001 (*Note*: The 'Western Offshoots' are defined as the United States, Canada, Australia, and New Zealand. 'Asia' excludes Japan. *Source*: Maddison (2003, Table 8c))

Growth in income per capita is thus an increasing function of productivity growth and a decreasing function of population growth. The negative effect of population growth reflects the fact that land is a fixed factor: when the size of the population increases, there is less land for each person to work with, which lowers income per capita.

To turn the growth Eq. (2) into a theory of stagnation, one needs to specify how productivity A_t and population N_t evolve over time. Assume for now that productivity growth is constant, $\gamma(A_t) = \overline{\gamma}_A$. The main assumption underlying the Malthusian theory of stagnation is that population growth is an increasing function of income per capita y_t:

$$\gamma(N_t) = f(y_t), \tag{3}$$

where $f'(y_t) > 0$. A number of different justifications can be given for this relationship. One possibility is that children enter the utility function of parents as normal goods. A rise in income would then increase the demand for children, leading to higher population growth. Alternatively, the mechanism could also work through mortality. If higher income leads to better nutrition and, as a consequence, lower mortality rates, a positive relationship between income per capita and population growth follows. As an empirical matter, the assumption of a positive relationship appears to fit the experience of most pre-industrial economies rather well.

Using (3), the growth Eq. (2) reads:

$$\gamma(y_t) = \overline{\gamma}_A - (1-\alpha)f(y_t). \tag{4}$$

According to this equation, the *growth rate* of income per capita is a decreasing function of its *level.* If the detrimental effect of population growth is sufficiently strong, this mechanism leads to stagnation as the only possible long-run outcome. In a country where income per capita is initially rising, population growth will accelerate until it fully offsets productivity growth, $(1-\alpha)f(y_t) = \overline{y}_A$, resulting in stagnation.

The Malthusian model is remarkably successful in terms of explaining economic growth (or the lack thereof) until Industrial Revolution. However, we now know that ultimately many countries managed to escape from the Malthusian trap. In these countries, living standards today are far superior to what almost any human alive before 1800 could have experienced. How can this drastic change in the economic fate of countries be explained?

Endogenous Population Growth

Given the growth Eq. (4), one scenario that could lead to a growth takeoff is a reversal of the income–population feedback. If the positive relationship described by the equation $\gamma(N_t) = f(y_t)$ breaks down, and subsequent population growth

is low, growth will ensue. Consider, for example, the case where population growth ceases altogether, $\gamma(N_t) = 0$. According to Eq. (2), growth in output per capita is then equal to productivity growth. Thus, as long as productivity keeps increasing, income per capita will grow indefinitely.

Historically, the Malthusian relationship between income and population growth did indeed break down in every single country that experienced a growth take-off. In a pattern known as the *demographic transition,* the high fertility and mortality rates of the pre-industrial era gave way to a new regime in which fertility, mortality, and population growth are low. In modern data, the relationship between income per capita and population growth is negative (both in a cross section of countries and in the time series for most rich countries), which is the opposite of what the Malthusian model assumes.

Figure 2 illustrates the demographic transition by comparing population growth in western Europe (the first region to experience a take-off) with Asia and Africa (the regions that stagnated the longest). In western Europe, population growth reached a peak at the end of the nineteenth century and has been declining since, despite rapid growth in income per capita. In Asia and Africa, in contrast, population growth has accelerated since the mid-nineteenth century, and is now much higher than in western Europe.

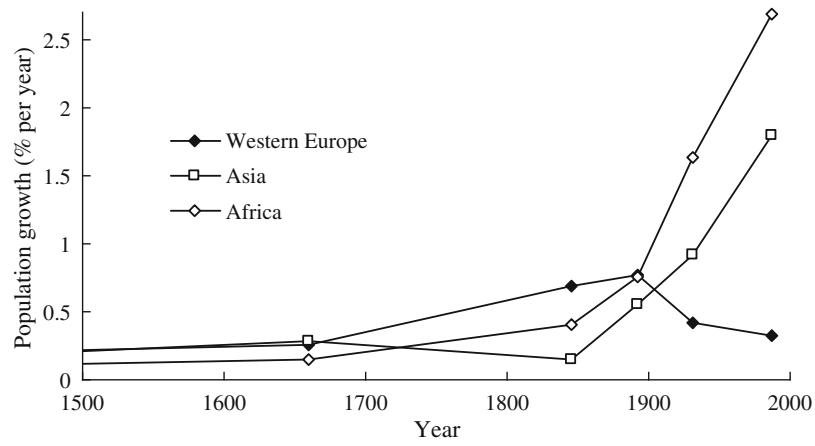
A number of authors have developed theories that integrate models of economic growth and the demographic transition to explain growth take-offs. In this literature, fertility decline is usually interpreted as a substitution of child 'quantity' (a large number of children) by child 'quality' (fewer children in which parents invest in terms of education or human capital). As an example of a model capturing this tradeoff, consider the decision problem of a parent with preferences

$$u(c, n, h) = (1 - \beta)\log(c) + \beta[\log(n) + \gamma\log(h)]$$

over consumption c, the number of children n, and the children's human capital h, where $\beta > 0$ and $0 < \gamma < 1$. The parent has to spend fraction φ of its time to raise each child, and can choose to spend an additional per-child fraction e on educating the children. The total child-rearing time is then given by $(\varphi + e)n$, and the budget constraint for the parent is $c = (1-(\varphi + e)n)wH$, where H is the parent's human capital, w is the wage per unit of human capital, and the time endowment is normalized to one. A child's human capital depends on the parent's human capital H and education time e:

$$h = 1 + \mu He,$$

where μ is the productivity of the education technology. Notice that a child receives at least one unit of human capital even if education e equals zero, which represents basic productive skills (such as physical strength) that do not rely on



Growth Take-Offs, Fig. 2 The evolution of population growth across world regions, years 1500–2001 (*Note:* 'Asia' excludes Japan. *Source*: Maddison (2003, Table 8a))

education. Lastly, the parent also has to observe a subsistence consumption constraint, $c \geq \overline{c}$, where $\overline{c}$ is the minimum amount of consumption required for survival.

In this model, the relationship between income and fertility depends on whether the optimal choices for education and consumption are at a corner. Assume that, initially, the wage w and the education productivity μ are so low that the subsistence constraint is binding and the parent chooses zero education ($e = 0$). The number of children is then constrained by the need to earn at least $\overline{c}$ units of consumption:

$$n = \frac{1}{\varphi}\left(1 - \frac{\overline{c}}{wH}\right).$$

Under this regime, the relationship between income wH and fertility n is positive, as assumed by the Malthusian model.

The outcome changes substantially if, through an increase in the wage w and the education productivity μ, the economy enters a regime where the subsistence constraint is no longer binding, and education is positive: $e > 0$. Under this regime, parents spend a fixed fraction of their time on child rearing. The balance between child quality and quantity depends on parental human capital H. The optimal decision rules are:

$$n = \frac{\beta}{\varphi + e} \text{and} \tag{5}$$

$$e = \frac{1}{1-\gamma}\left(\gamma\varphi - \frac{1}{\mu H}\right). \tag{6}$$

Equation (5) captures the trade-off between child quality and quantity: the number of children is a decreasing function of education e. Intuitively, investing a lot in each child renders children expensive, which reduces demand. Education e, in turn, depends positively on parental human capital H. An increase in income per capita (through a rise in H) therefore lowers fertility n, the opposite of the Malthusian assumption.

Given these results, an escape from the Malthusian trap is possible if some change in the economy generates increased investment in child quality. The literature proposes different candidates for the underlying cause of such an event. In Galor and Weil (2000), the take-off is ultimately a consequence of technological progress. Accelerating productivity growth increases the return to education (the parameter μ in the model outlined above), which eventually triggers the quantity–quality substitution and the growth take-off. Galor and Moav (2002), in contrast, suggest that evolving parental preferences (through an increase in the parameter γ) are the driving force behind fertility decline. Yet other authors have emphasized the role of declining mortality rates (Boucekkine et al. 2002; Cervellati and Sunde 2005; Doepke 2005; Kalemli-Ozcan 2002; Lagerlöf 2003a; Soares 2005), increasing female labour-force participation (Galor and Weil 1996; Lagerlöf 2003b), changes in the provision of old-age security (Boldrin and Jones 2002), changes in child-labour and education laws (Doepke and Zilibotti 2005), and the introduction of skill-intensive production technologies that raise the return to education (Doepke 2004).

Structural Change

Apart from endogenous population growth, the Malthusian model also relies on the presence of the fixed factor of land to generate stagnation. A second potential trigger for a growth take-off is therefore *structural change* that decreases the role of land. In pre-industrial economies, agriculture was the main mode of production. In contrast, in modern industrial economies the share of agriculture in output is small, and consequently land is less important. Translated into the growth Eq. (4), structural change amounts to a shift in the parameter α. In particular, an increase in α lowers the detrimental effect of population growth on income per capita. In the limit case of $\alpha = 1$, income per capita is independent of the size or growth rate of the population, and is solely driven by productivity growth.

In Hansen and Prescott (2002), a decline of the role of land is generated endogenously in an environment where two competing technologies can be used for production. (Related contributions

include Matsuyama 1992; Laitner 2000; Kögel and Prskawetz 2001; Gollin et al. 2002; Ngai 2004). In addition to the production function (1) above, an 'industrial', constant-returns technology is also available:

$$Y_t^I = A_t^I N_t^I,$$

where Y_t^I is industrial output, Y_t^I is productivity, and N_t^I is the amount of labour employed in the industrial sector. Productivity A_t^I is assumed to grow at a constant rate. The total amount of labour is allocated optimally between the traditional sector and the industrial sector. Given the linear production technology, output per worker in the modern sector is given by A_t^I. Early in development, when A_t^I is still low, it is optimal to allocate all workers to the traditional sector. During this phase the economy behaves just like a Malthusian economy where the modern technology does not exist at all.

Ultimately, however, the modern technology becomes sufficiently productive to be introduced. If w_M is the (constant) marginal product of a worker in the Malthusian regime, the technology will be introduced once $A_t^I > w_M$. From this point on, population growth no longer affects output per worker, since land is not used in the industrial sector. Output per worker therefore starts to grow at the rate of technological progress. Viewed through the lens of the Hansen–Prescott model, what initially appears as a structural break in economic history is merely the outcome of an optimal sectoral allocation decision in an otherwise stable economic environment.

Endogenous Technological Progress

Starting once again from the growth Eq. (4), a third potential trigger for a growth takeoff is a sustained increase in productivity growth that is large enough to 'outrun' population growth. Clearly, population growth cannot increase indefinitely, as there are physiological constraints on child bearing. Let $\overline{\gamma}_N$ be an upper bound for population growth that cannot be exceeded for biological reasons. If now productivity growth satisfies

$$\gamma_A > (1-\alpha)\overline{\gamma}_N,$$

even at maximum population growth the detrimental effect of increasing population density does not suffice to negate productivity improvements, and improving living standards ensue.

A potential cause for accelerating productivity growth is scale effects in the production of ideas. An increase in world population implies that there are more people who might invent new, productive technologies. An increase in world population should therefore imply an acceleration of productivity growth. Scale effects of this kind underlie the takeoff models of Kremer (1993), Jones (2001), and Tamura (2002).

Conclusions

The three potential triggers for a growth take-off presented here should not be regarded as mutually exclusive alternatives, but rather as complementary explanations for a joint phenomenon. From an empirical perspective, there is little doubt that all three explanations are relevant: every country that underwent a growth takeoff also experienced a demographic transition, a sectoral shift from agriculture to industry and services, and an acceleration of productivity growth. Reflecting these observations, many papers in the literature already incorporate more than one of the mechanisms. For example, a number of authors propose models where accelerating endogenous productivity growth triggers a fertility transition. This is true, for example, of the seminal paper of Galor and Weil (2000) and, in a framework driven by human-capital externalities, for de la Croix and Doepke (2003). Similarly, Greenwood and Seshadri (2002) and Doepke (2004) integrate models of structural change with theories of fertility decline.

Building on the different mechanisms behind growth take-offs that have been proposed in recent years, a major challenge for future research is to understand why in many countries

these mechanisms fail to work to the present day. Conceivably, a better understanding of the mechanisms that allowed some countries to overcome economic stagnation two centuries ago might help us learn how the same feat could be accomplished in poverty-stricken developing countries today.

See Also

- ▶ Demographic Transition
- ▶ Economic Growth in the Very Long Run
- ▶ Industrial Revolution
- ▶ Malthus, Thomas Robert (1766–1834)
- ▶ Population and Agricultural Growth
- ▶ Poverty Traps

Bibliography

Boldrin, M., and L.E. Jones. 2002. Mortality, fertility, and saving in a Malthusian economy. *Review of Economic Dynamics* 5: 775–814.

Boucekkine, R., D. de la Croix, and O. Licandro. 2002. Vintage human capital, demographic trends, and growth. *Journal of Economic Theory* 104: 340–375.

Cervellati, M., and U. Sunde. 2005. Human capital formation, life expectancy and the process of development. *American Economic Review* 95: 1653–1672.

de la Croix, D., and M. Doepke. 2003. Inequality and growth: Why differential fertility matters. *American Economic Review* 93: 1091–1113.

Doepke, M. 2004. Accounting for fertility decline during the transition to growth. *Journal of Economic Growth* 9: 347–383.

Doepke, M. 2005. Child mortality and fertility decline: Does the Barro–Becker model fit the facts? *Journal of Population Economics* 18: 337–366.

Doepke, M., and F. Zilibotti. 2005. The macroeconomics of child labor regulation. *American Economic Review* 95: 1492–1524.

Galor, O., and O. Moav. 2002. Natural selection and the origin of economic growth. *Quarterly Journal of Economics* 117: 1133–1191.

Galor, O., and D.N. Weil. 1996. The gender gap, fertility, and growth. *American Economic Review* 86: 374–387.

Galor, O., and D.N. Weil. 2000. Population, technology, and growth: From Malthusian stagnation to the demographic transition and beyond. *American Economic Review* 90: 806–828.

Gollin, D., S. Parente, and R. Rogerson. 2002. The role of agriculture in development. *American Economic Review* 92: 160–164.

Greenwood, J., and A. Seshadri. 2002. The U.S. demographic transition. *American Economic Review* 92: 153–159.

Hansen, G.D., and E.C. Prescott. 2002. Malthus to solow. *American Economic Review* 92: 1205–1217.

Jones, C.I. 2001. Was an Industrial Revolution inevitable? Economic growth over the very long run. *Advances in Macroeconomics* 1(2), Article 1. Online. Available at http://www.bepress.com/bejm/advances/vol1/iss2/art1. Accessed 5 Oct 2006.

Kalemli-Ozcan, S. 2002. Does the mortality decline promote economic growth? *Journal of Economic Growth* 7: 411–439.

Kögel, T., and A. Prskawetz. 2001. Agricultural productivity growth and the escape from the Malthusian trap. *Journal of Economic Growth* 6: 337–357.

Kremer, M. 1993. Population growth and technological change: One million B.C. To 1900. *Quarterly Journal of Economics* 108: 681–716.

Lagerlöf, N. 2003a. From Malthus to modern growth: Can epidemics explain the three regimes? *International Economic Review* 44: 755–777.

Lagerlöf, N. 2003b. Gender equality and long-run growth. *Journal of Economic Growth* 8: 403–426.

Laitner, J. 2000. Structural change and economic growth. *Review of Economic Studies* 67: 545–561.

Maddison, A. 2003. *The world economy: Historical statistics*. Paris: OECD.

Malthus, T.R. 1798. *Essay on the principle of population*. Harmondsworth: Penguin, 1970.

Matsuyama, K. 1992. Agricultural productivity, comparative advantage, and economic growth. *Journal of Economic Theory* 58: 317–334.

Ngai, R. 2004. Barriers and the transition to modern growth. *Journal of Monetary Economics* 51: 1353–1383.

Soares, R. 2005. Mortality reductions, educational attainment, and fertility choice. *American Economic Review* 95: 580–601.

Tamura, R. 2002. Human capital and the switch from agriculture to industry. *Journal of Economic Dynamics and Control* 27: 207–242.